VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Validity generalization revisited

Jansen, Paul; Roe, Robert A.; Vijn, P.; Algera, J.A.

1986

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Jansen, P., Roe, R. A., Vijn, P., & Algera, J. A. (1986). *Validity generalization revisited*. Delft University Press.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 05. May. 2023

VALIDITY GENERALIZATION REVISITED

Paul G.W. Jansen, Robert A. Roe,
Pieter Vijn, Jen A. Algera

2224 4442

Delft University Press, 1986

1594486

C10091
25772

VERVALLEN

VALIDITY GENERALIZATION REVISITED

BIBLIOTHEEK TU Delft
P 2224 4442

C 1580883

7503

Paul G.W. Jansen
The Netherlands Postal and Telecommunications Services, The Hague

Robert A. Roe
Delft University of Technology, Delft

Pieter Vijn
Wunderman International B.V. Amsterdam

Jen A. Algera
Free University, Amsterdam

VALIDITY GENERALIZATION REVISITED

Paul G.W. Jansen
Robert A. Roe
Pieter Vijn
Jen A. Algera

2224 4442

87

012103

SCHOTTENVELD 2
2600 *DELFT* NEDERLAND

Delft University Press / 1986

Published and distributed by:

Delft University Press
Stevinweg 1
2628 CN Delft
The Netherlands
Telephone: (015)-78 32 54

By order of:
Vakgroep Techniek, Arbeid en Organisatie,
Sectie Arbeids- en Organisatiepsychologie,
Kanaalweg 2B,
2628 EB Delft,
Tel. (015) 78 37 20.

CIP—DATA KONINKLIJKE BIBLIOTHEEK, THE HAGUE

ISBN 90-6275-295-2

Copyright 1986 by Delft University Press, The Netherlands.
All rights reserved.
No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electric or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from Delft University Press.

Printed in The Netherlands

Contents

Abstract

The Schmidt et al. procedure for generalizing validities obtained in research on predictor-criterion relationships in personnel selection, is critically discussed. The three components of the method, viz. (1) a procedure for compiling and classifying validity data, (2) a procedure for testing the homogeneity of a set of data classified as homogeneous, and (3) a procedure for making generalizations on account of such data, are scrutinized at both a theoretical-conceptual, and a methodological-statistical level. Generally, it appears that the procedure is liable to improvements at both these levels. Specifically, when the procedure is remodeled in a proper Bayesian sense, it appears to be not robust to applications on data that violate the assumptions of Schmidt et al.'s underlying rather confined psychometric model of validity generalization. Finally, suggestions are made to improve Schmidt et al.'s validity generalization procedure.

VALIDITY GENERALIZATION REVISITFD

1. Introduction

In developing a personnel selection procedure one needs information about the validity of the predictor instruments against the criteria at hand. Such information can be obtained in at least two different ways: validities can be estimated from (1) an empirical study on the specific case, performed ad hoc, or (2) accumulated results of previously conducted studies on similar cases. The choice between these two options is the subject of some controversy. Traditionally, researchers have preferred the first approach, following Ghiselli (1966, 1973) who compiled and analyzed validity data from a large number of studies and found that validities observed for tests carrying the same name varied too much to allow reliable generalizations. In their view empirical validation is needed because a unique set of factors determines validity in every single case. The second approach is advocated by Schmidt et al. In their view the greater part of the observed validity variance can be attributed to factors like small sample size, criterion unreliability, restriction of range etc., called 'artifacts' by them (Schmidt et al., 1979). They hold that when these factors are taken into account validity generalization is perfectly feasible and ad hoc validation studies are no longer required (Schmidt & Hunter, 1977; Schmidt et al., 1979, 1980, 1981a, b, 1982; Pearlman et al., 1980).

The basis of Schmidt et al. 's position lies in a series of analyses of published validity data, performed with the help of a 'bayesian validity generalization method', specifically devised for this purpose (Schmidt & Hunter, 1977) and revised a number of times (Schmidt et al., 1979; Pearlman et al., 1980). This method has three components:

1) a procedure for compiling and classifying observed validity data,
2) a procedure for evaluating the heterogeneity of a given set of data at the level of underlying theoretical constructs,
3) a procedure for making generalizations from such sets of observed validity data.

Schmidt et al. have advocated the use of the method as a general tool for theoretical research on situational specificity and moderator phenomena (Schmidt & Hunter, 1977; Schmidt et al., 1981b). On the basis of their studies, they have arrived at far reaching conclusions on the true validities of ability tests and all kinds of methodological issues in personnel selection.

For instance, Schmidt and Hunter (1981) state: 'Professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs in all settings' (p. 1128), 'there is no factual basis for requiring a validity study in each situation' (p. 1133), 'there is no empirical basis for requiring separate validity studies for each job: tests can be validated at the level of

job families' (p. 1133), 'these findings effectively show the theory of situational specificity to be false' (p. 1132). In the article by Schmidt et al. (1981b) the well-established multidimensionality of criteria (Schmidt, 1976) is denied, as is the role of factors like organizational climate, management philosophy or leadership style, geographical location, changes in technology, product or jobs over time, age, socio-economic status, and applicant pool composition as moderators of test validities (p. 175-176).

In this study, Schmidt et al.'s validity generalization method is critically examined, both at a conceptual (section 2) and a psychometric/statistical (section 3) level. Some points of criticism have been presented before (Algera et al., 1984; Roe, 1984; Roe et al., 1983 a, b); here they will be presented in more detail. Suggestions will be made for improving the method (sections 4 and 5).

2. Conceptual issues/problems of definition

We set out with a discussion of some conceptual issues, relating to the definitions of predictive validity, validity generalization, situation, and situational specificity. Schmidt et al., have not been very explicit on these issues. We feel, however, that a thorough examination is crucial to a proper understanding of the problems met with regard to validity generalization.

2.1. Predictive validity

In modern selection theory, the notion of predictive validity has a two-fold meaning. On the one hand it refers to the linear correlation between a predictor variable X_h and a criterion variable Y_i observed in a sample of applicants P_j to a given job, while on the other hand it refers to the linear correlation between a predictor construct ξ and a criterion construct η, of which X_h and Y_i are operationalizations, in a population of applicants Π from which the sample P_j is drawn. These two meanings should be clearly distinguished. The first type of validity requires specific tests and criterion instruments which are interrelated within a specific sample. This validity will be called observed validity. With the second type, test and criterion variables are defined at the construct level, e.g. 'verbal reasoning' and 'quality of performance', and their relationship in the population is a hypothetical one. This type of validity will be denoted as theoretical validity. Because of selection on the predictor X_i, the sample on which the validity is actually computed mostly will be restricted in range. If this is the case, the sample will be denoted as p_j. However, since range restriction can, in principle, be corrected for, the observed validity usually will be written as $r_{x_h y_i P_j}$ in the sequel.

It should be noted that both types of validity have three defining terms, or 'referents': predictor, criterion, and sample, resp. predictor construct, criterion construct,

and population. These must be known in order to determine or interprete validity coefficients. For this reason we will specify them from now on, writing $r_{x_h y_i P_j}$ for the observed validity and $\rho_{\xi\eta\Pi}$ for the theorical validity.

Implications of the foregoing are that within a given sample a predictor may show different validities for different criteria, just as different predictors may show different validities for the same criterion, and further that the validity of a given predictor for a given criterion may be dependent on the nature of the sample. E.g. the validity of a spatial ability test for the prediction of an accident criterion may by different for a sample of rural high school drivers, adult suburban female drivers, metropolitan cab drivers, and long distance truck drivers.

While predictors and predictor constructs can be directly defined by referring to psychological instruments or theory, the other defining terms cannot. The criterion and the sample (or criterion construct and population) refer to a certain job within a specific company (or a job type within a sector of industry). Productivity, quality, turnover, accident criteria can only be measured, and measures can only be interpreted, if content and context of the jobs are know.

In fact, more defining terms might be distinguished, like for instance the time interval between the moments of predictor and criterion measurement, the conditions of measurement (such as: the way in which the tests are

administered, criterion ratings are generated: by the same or different persons, etc.), and the specific working conditions (enabling individual characteristics more or less to influence job performance). For the sake of simplicity we leave these out of account however.

The two types of validity can be related by introducing assumptions on the relationships between the referents, i.e. X and ξ, Y and η, P and Π. In this way, specific psychometric models may be set up which can serve as a basis for generalization. Below, we present the model as employed by Schmidt et al., and an alternative model.

A single-test-single-criterion-model

The psychometric model adopted by Schmidt et al. is based on classical test theory (Schmidt et al., 1982, p. 836). In it the test true score T_x takes the place of the predictor construct ξ, while the criterion true score T_y stands for the criterion construct η. In this way the constructs have a narrow meaning: they cover only one test and one criterion instrument. As a result the theoretical validity has a limited meaning also. It is the population correlation between the true score components of the specific predictor and the criterion instruments (see figure 1). Schmidt et al. denote it as 'true validity'.

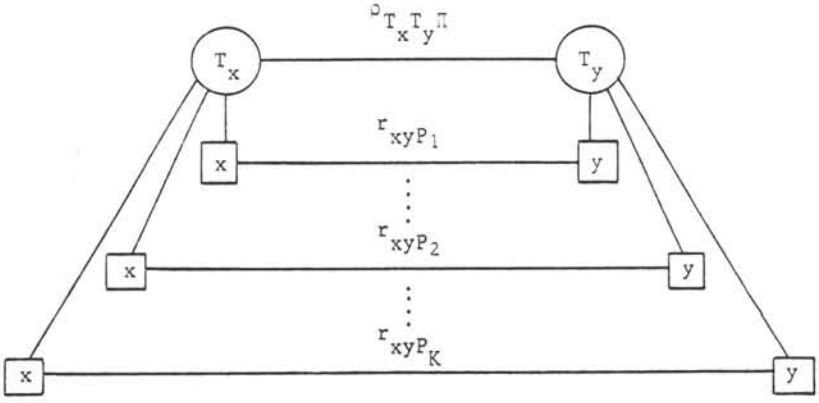

Figure 1 : STSC model for validity generalization

The true scores T_{x_m} of individuals 1 ... m ... N represent the parts of the test scores X_m that can be measured with perfect reliability. The relationship between the observed score and the true score of subject m is described by the following basic equation from classical test theory:

$$X_m = T_{x_m} + E_{x_m} . \qquad (1)$$

The same holds for the criteria:

$$Y_m = T_{y_m} + E_{y_m} . \qquad (2)$$

The error components E_{x_m} and E_{y_m} of (1) and (2) are assumed uncorrelated. The theoretical validity is conceived as the true-score correlation

$$\rho_{T_x T_y \Pi} = \frac{\text{cov}(T_x, T_y)}{\sigma(T_x) \cdot \sigma(T_y)} \quad , \qquad (3)$$

within a given population Π. In 'empirical terms' it is the validity of test X against criterion Y that would be observed if

- sample size were infinite (i.e. sampling error were zero),
- criterion reliability were perfect,
- test reliability were perfect,
- range restriction were absent,

(Schmidt et al., 1979, p. 266).

The true validity $\rho_{T_x T_y \Pi}$ can be estimated from the observed, range restricted validities r_{xyp} in different ways. Most straigthforward would be to correct every single r_{xyp} for attenuation and range restriction, and take the average of these corrected validities as an estimate for $\rho_{T_x T_y \Pi}$. In order to make corrections on the individual validities specific data would be required. Since these were not available, Schmidt et al. first computed the sample-size weighted mean of the observed validities, and next corrected this mean observed validity for attenuation and range restriction, using assumed average values for test reliability (e.g. .80), criterion reliability (e.g. .60),

and range restriction (e.g. the ratio of the restricted standard deviation to the unrestricted standard deviation could be assumed .60). This corrected mean observed validity was taken as an estimate of $\rho_{T_x T_y \Pi}$ (cf. Pearlman et al., 1980, pp. 402-406; see also appendix A).

The model adopted by Schmidt et al. could be qualified as a single-test-single-criterion-model (STSC-model). It allows generalizations from a series of observed validities of a given test-criterion combination to future observations of the same validity in other samples from the same population. However, Schmidt et al. have also and more frequently used the model for generalizing from validities with varying test, criterion, and sample referents to future validities of any type.

Multiple-test-multiple-criterion models

In our view, generalizations of the type Schmidt et al. aim at require another psychometric model. It would have to include a predictor construct ξ which relates to multiple tests $X_1 \ldots X_h \ldots X_L$ and a criterion construct η which relates to multiple criteria $Y_1 \ldots Y_i \ldots Y_M$. A basic model that satisfies this requirement is presented below.

We assume the tests X_h to be 'congeneric' (Lord & Novick, 1968): they share a latent trait component ξ. Further we assume that the relationships between ξ and its indicators X_h follow a linear model:

$$X_h = \beta_h \xi + \delta_h, \qquad (4)$$

in which β_h is the loading of test X_h on factor ξ and δ_h is the residual part of X_h, that cannot be explained by ξ. In the same way:

$$Y_i = \gamma_i \eta + \varepsilon_i, \qquad (5)$$

in which γ_i is the loading of the criterion Y_i on factor η an ε_i is the residual part of Y. Again, δ_h and ε_i are assumed uncorrelated. The theoretical validity then is defined as the correlation between ξ and η:

$$\rho_{\xi\eta\Pi} = \frac{\text{cov}\ (\xi,\eta)}{\sigma(\xi)\ .\ \sigma(\eta)}\ , \qquad (6)$$

within the given population. It relates to the observed validities by a set of formulas to be given in section 4.

In order to estimate $\rho_{\xi\eta\Pi}$ from a number of observed validities $r_{x_h y_i P_j}$, the regression coefficients β_h and γ_i should be estimated first. This could be done, for instance, by means of the LISREL-procedure (cf. Jöreskog, 1973, 1974, 1978; Jöreskog & Sörbom, 1978), provided sample size is large enough and certain assumptions are met.

This model (see figure 2) is a <u>multiple-test-multiple-criterion-model</u> (MTMC-model). It allows generalizations from validity data on different but congeneric tests and criterion instruments to future validities of tests and

criteria from the same domain, to be observed in samples from the same population.

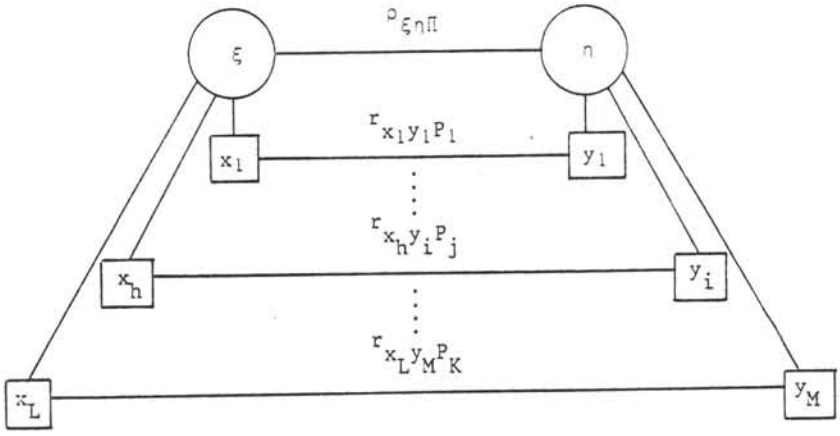

Figure 2 : Basic MTMC model for validity generalization

The STSC and MTMC models can be related by classical test theory. Assuming that every test contains, apart from its common factor ξ, a specific component α_h that is stable and can be measured therefore with perfect reliability, and a residual component E_{x_h} that is even unstable within the same test, we can rewrite (4) as

$$X_h = \beta_h\xi + (\alpha_h + E_{x_h}), \qquad (7)$$

in which by definition $(\beta_h\xi + \alpha_h)$ is equivalent to T_{x_h} of

formula (1), the part that can be reliably measured. Analogously:

$$Y_i = \gamma_i \eta + (\theta_i + E_{y_i}), \qquad (8)$$

in which $(\gamma_i \eta + \theta_i)$ can be recognized as T_{y_i} from formula (2). Again, X_h and θ_i are assumed uncorrelated.

Equations (7) and (8) define an extended MTMC-model (see figure 3), which embraces two types of theoretical validity:

- Schmidt et al.'s true validity, the correlation between the true score components of the STSC-model;
- construct validity in a more general sense, i.e. the correlation between the construct terms of the MTMC-model.

Obviously, these theoretical validities are, generally, not identical.

The extended MTMC-model reduces to the STSC -model when for different predictors and criteria the loadings β, γ, and the error terms E_x and E_y are equal. This illustrates the position of the extended MTMC-model an an intermediate between the STSC-model and the basic MTMC-model. As the extended MTMC-model only serves to clarify the relationships between these two models, it will not be discussed any further.

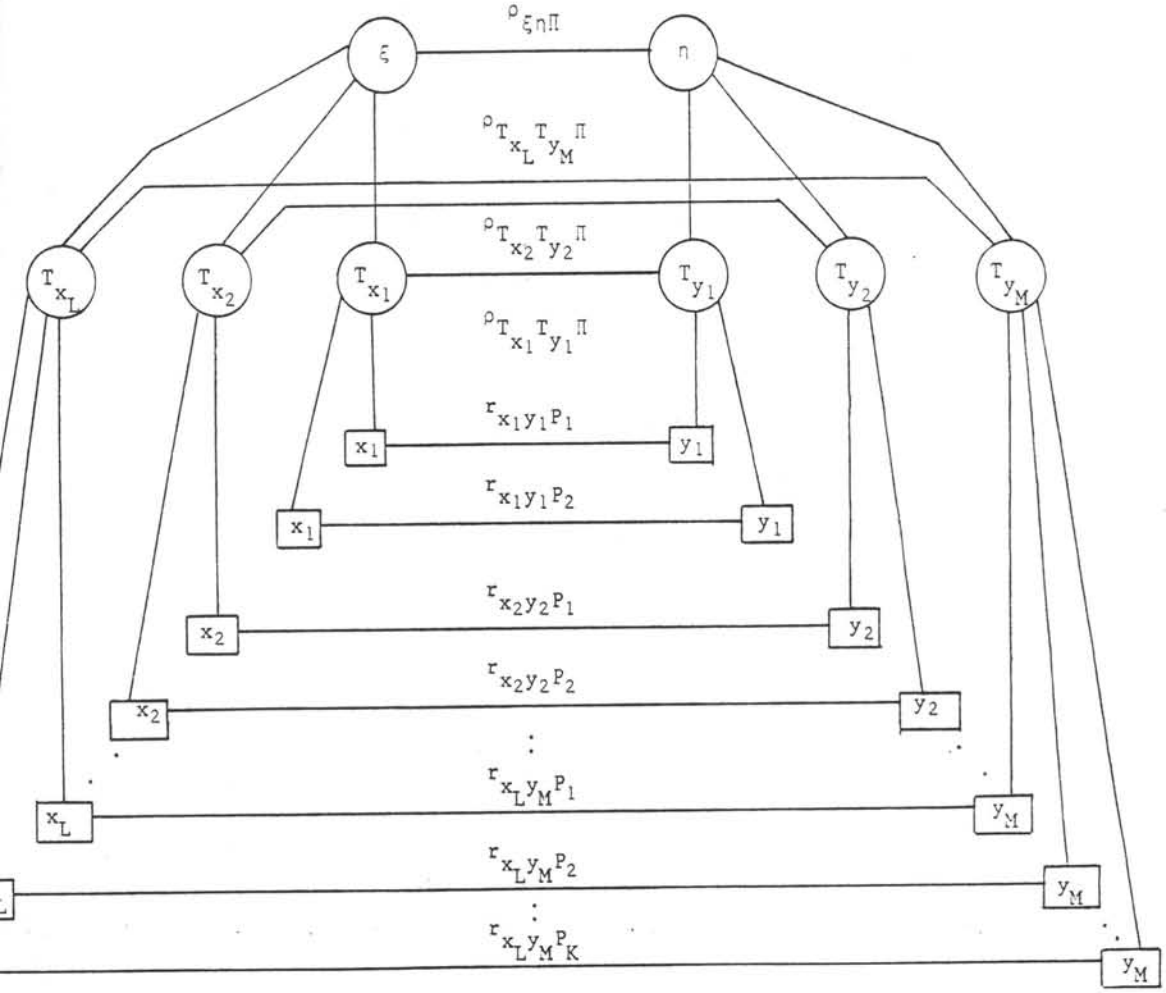

Figure 3 : Extended MTMC-model for validity generalization

2.2. Validity generalization

Since its first use by Lawshe (1952) the term validity generalization has become a label for an inferential process: the estimation of the numerical value of theoretical validity from a series of observed validities. This estimation can be produced either in an judgmental way, or by statistical procedures. In the STSC-model the estimate $\hat{\rho}_{T_x T_y \Pi}$ is derived from data about a single test and a single criterion, coming from different samples. Within the framework of the MTMC-model an estimate $\hat{\rho}_{\xi\eta\Pi}$ is derived using several types of r-data (see figure 4). In this case observed validities may come from different predictors, criteria and samples, provided that these relate to the predictor construct ξ, criterion construct η (and applicant population Π).

There is a deductive counterpart to this inductive process. From a given ρ (or estimated $\hat{\rho}$) an estimate of future r's can be obtained again, either judgmentally or statistically, although the latter approach is more usual and has some advantages. Given certain statistical assumptions, both interval and point estimates of a future $r_{x_h y_i P_j}$ can be derived by the model used in the inductive phase. We feel that this second phase, which is not explicitly mentioned by Schmidt et al., should be considered as an integral part of validity generalization. Without it, validity generalization would be of theoretical value only, and irrelevant to the practice of personnel selection.

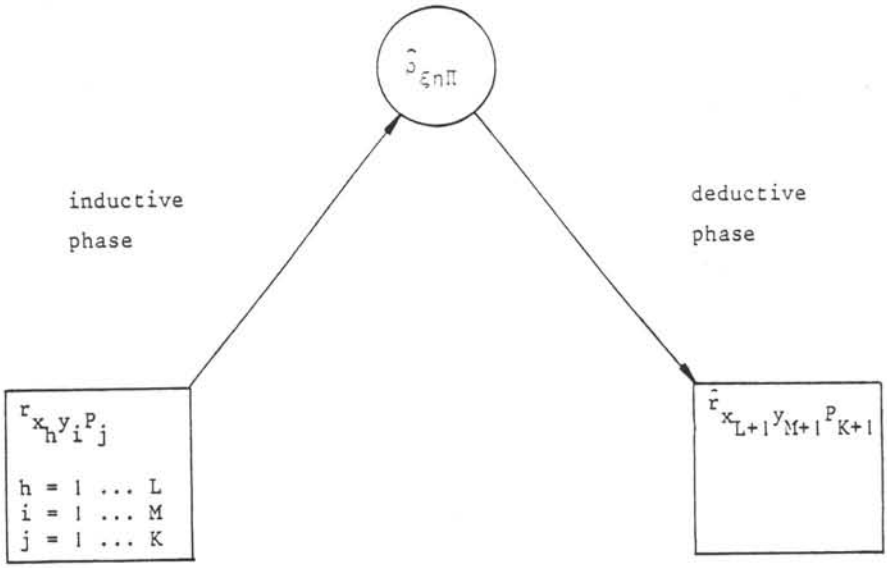

Figure 4 : A two-phase view of validity generalization by the MTMC model

2.3. Situation and situational specificity

Following modern personality theory (e.g. Magnusson, 1981; Ekehammer, 1974) one would be inclined to define 'situation' as the whole of those factors that, without being tied in any way to the individual, help to determine his work behavior. While some of these factors will operate in additive manner, some others may interact with personality traits, thus acting as moderators of the

relationship between individual characteristics and work behavior. Predictive validities of tests measuring individual characteristics would vary under the control of such factors and hence show 'situational specificity'.

In the classical literature on personnel selection the two concepts have a broader scope. Ghiselli (1959) called test validities situationally specific when he noted that observed values, stemming from samples of applicants to the same jobs and relating to the same predictors, showed considerable differences, falling outside the range to be expected on the basis of sampling error. Referring to studies with an N of at least 100, he gives the following examples: 'For the 71 reports I was able to find for intelligence tests applied to general clerks (the validation being against proficiency criteria) the range in validity coefficients was from about -.40 to 0.80. The middle 50 per cent of the coefficients covered a range of 0.50 correlation points. For 99 reports of spatial relations tests (validity against proficiency criteria) for machine-tenders, the validity coefficients ranged from -.55 to +.65, with the middle 50 per cent of the coefficients covering a range of .35 correlation points', which leads him to conclude that '... the variation among the reported validity coefficients for a given test applied to workers in a given job cannot be entirely explained on the basis of sampling error from some population avarage' (Ghiselli, 1959, p. 398). Other authors (e.g. Lawshe & Balma, 1966), have conceived test validities

as situation specific by definition, because they bear on relationships of tests with specific job criteria established in specific samples.

So it seems that for classical theorists 'situational specificity' has been roughly equivalent to: depencency on other factors than those used in the definition of validity (i.e. the referents), at their time being the predictor instrument (test) and a job title. In that case 'situation' meant: the whole of these other factors, including the working conditions relevant for performance, the specific nature of criteria, sample size and composition, the time interval between predictor and criterion measurement, measurement characteristics like the manner of test administration, scoring accuracy, etc.

Generally, three kinds of factors are involved:

1. behavioral determinants from 'outside the individual', such as restrictive working conditions or motivational contingencies;
2. factors having to do with the aspects of work behavior that are being considered, like the content of the job, the dimensionality of performance, and the nature of the criterion instruments (including criterion contamination and deficiency);
3. factors that relate to the research design adopted for the validation study, like the time interval between measurements, unwarranted unreliability, restriction of range, etc.

Schmidt et al. have taken another approach. In their view, when the hypothesis of situational specificity holds, differences in validity for a given test-job combination are due to differences in the factor structure of job performance (Schmidt & Hunter, 1977; Schmidt et al., 1979). Thus, their definition of 'situation' is limited to the second category, leaving out the first as well as the third. In fact, some of the factors from the third category are labelled 'artifacts' by them.

We feel that this nomenclature may bring confusion. For instance, in the case of an equal factor structure of global criteria and great artifactual effects on observed validities, Schmidt et al. would declare situational specificity to be absent, while other researchers would state that situational specificity was present.

We prefer to follow the traditional approach, and use the term situational specificity as a summary label for the phenomenon that observed validities are dependent on factors not considered in defining the validity concept. At the same time however, we suggest that a clear distinction be made between the three sources of situational specificity: external behavioral determinants, job and criterion characteristics, and research design parameters (including 'artifacts'). Knowledge about the degree to which factors from these sources influence validities is highly desirable, because it may improve our understanding of work behavior, work organization, and selection research methodology.

A final point to note, is that the approach that we have chosen makes the concepts 'situation' and 'situational specificity' dependent on the definition of validity. More complete definitions, involving three or even more referents, will restrict the content of these concepts. When all moderating determinants, job-criterion factors and research design parameters would be included in the definition of validity, 'situation' would become equivalent to sample size, and 'situational specificity' to sampling error for observed validities, while at the level of theoretical validities both terms would have no meaning whatsoever.

3. Discussion of the Schmidt et al. method

The three components of Schmidt et al.'s method as briefly referred to in the introduction, are:

1. Compilation-classification:

   Previously observed validity data are collected, and classified into more or less homogeneous sets on account of type of predictor test, type of job, and type of criterion measure. In principle, each data set in such a test-job-criterion class is to be analyzed separately.

2. Generalizability testing:

   The validities classified into one class in the previous step are evaluated on statistical homogeneity and/or a minimum level of validity, in order to establish generalizability. To this aim, the average and variance

of the 'residual distribution' of validities are computed, i.e. of the distribution that remains after the effects of 'artifacts' like sampling error, restriction of range, and attenuation have been statistically removed (see appendix A for a detailed presentation of this element of Schmidt et al.'s procedure).

3. Generalization:

Correcting the mean of the residual distribution from the previous step 'upward' on the basis of assumed average levels of criterion attenuation and range restriction, a point estimate of the true population validity is obtained. In some cases predictor attenuation is also corrected for. In a similar way the variance of the distribution of true validites is estimated. Assuming the latter distribution to be normal, a (e.g. 90%) lower bound estimate of the true validity can be established (see also appendix A).

It will be demonstrated below that all three components of the method as described by Schmidt and Hunter (1977), Schmidt et al. (1979), Pearlman et al. (1980), can be criticized, either on logical, methodological or statistical grounds.

3.1. Compilation-classification

In the first study (Schmidt & Hunter, 1977) the procedure is applied to four observed validity distributions

presented by Ghiselli (1966). These distributions, which contained both published and unpublished validity coefficients, are considered by Schmidt et al. to pertain to "similar" jobs and tests. They do not mention any explicit a priori rule for classifying validity data from different studies, but followed the crude classification scheme of Ghiselli.

This first study might be considered simply a demonstration of the Schmidt et al. generalization method leading to results that differ from those obtained by Ghiselli, without implications for the practice of personnel selection. When the focus is on the latter, some knowledge must be at hand about (at least) the test, job, and criterion types the validities in the analyses refer to. In subsequent studies therefore, schemes are presented for the classification of raw validities in different test types and/or job types and/or criterion types.

For instance, in the Schmidt et al. (1979) study some 3300 validity coefficients for various kinds of tests were located in the clerical area. Published as well as unpublished studies were included, also many older studies "squirreled away in dusty files" (p. 262). Tests were classified using a system derived from the classification schemes of Ghiselli (1966) and Dunnette (1972). For the classification of the clerical jobs the authors refer to "a slightly modified version of the Dictionary of Occupational Titles (DOT) classification system" (Schmidt et al., 1979;

p. 262). Criterion measures were indices of overall job performance or proficiency.

In the first large-scale application of their procedure for testing the hypothesis of no situational specificity, Pearlman et al. (1980) developed a data base of validity studies on clerical occupations. Ten general test types were established, most of which represent a construct or ability factor known from the literature. But also so-called "clerical aptitude" tests, motor ability tests, and performance tests were included, because of their relatively common use in clerical selection, even though they could be decomposed into more homogeneous constituent dimensions. The clerical jobs were grouped into five "true" DOT job family categories, one miscellaneous category and two additional categories developed to handle occupations that were not sufficiently specified in the original study to permit definitive classification, and samples representing two or more different clerical occupations. Criterion measures in this case were indices of job proficiency or training success.

To give an example, one cell of the test/job classification scheme in the Pearlman et al. study was verbal ability/DOT occupational groups 201-209. To get an impression of the broadness of this scheme one should note that the following jobs are included: 201 Secretaries, 202 Stenographers, 203 Typists and typewriting machine operators, 205 Interviewing clerks, 206 File clerks, 207

Duplicating machine operators and tenders, 208 Mailing and miscellaneous office machine operators, 209 Stenography, typing, filing, and related occupations, not elsewhere classified. With regard to test type, each cell contained different predictors. The verbal ability test type covered such predictors as reading comprehension, vocabulary, grammar, spelling, and sentence completion. Within this particular test/job cell 215 validity coefficients referring to criteria of overall job proficiency were compiled, from published and unpublished studies.

A general conclusion to be drawn from these and other studies is that classification rules vary over applications of validity generalization. Specifically with respect to test type, an evaluation of the different classification schemes is difficult, since Schmidt et al. generally refer to classes of tests rather than to specific predictors used in the actual studies. The varying schemes for classifying validities according to job type and criterion type are discussed in the next two sections.

Variability of classification rules: job type

"Job type" is a rather loosely defined category that may include different jobs in different companies, and may cover several criterion constructs and applicant populations. Even if one speaks of the same job in different settings (Schmidt & Hunter, 1981) there is no guarantee that the factorial composition of he criterion is identical. From the

literature on the measurement of task characteristics (see e.g. Roberts and Glick, 1981) it is known that individuals in the same job classifications are not always performing the same objective tasks. This underlines Ghiselli's (1966) statement that there exist differences in the nature of and requirements for nominally the same job in different organizations, and in the same organization from one time period to another. Schmidt and Kaplan (1971; p. 421) also refer to this last point: "Performance on some jobs has also been shown to be "dynamic", i.e., to change in apparent factorial composition over time ..... changes in organizational needs and goals can change the nature of the criteria of succes in individual jobs within the organization. Criterial dynamism is an important problem in industrial psychology, meriting much more research than has been devoted to it to date".

In personnel selection, the criterion (or criterion construct) refers to a certain job within specific organization. This implies that criterion measures (e.g. productivity, quality) can only be interpreted if the content and context of the job is known. What is needed here, is a taxonomy which takes both these content and context factors into account. In any case, test/criterion classification schemes would be required which are far more restricted than the broad schemes used by Schmidt et al. or Ghiselli.

The study of Colbert and Taylor (1978), in which jobs in the clerical domain were classified according to a very fine taxonomy, shows that in such a case, even when all jobs are localized in the same organization (an insurance company in this study), differential validity of different predictors for different jobs may be found (i.e. situationally specific validity).

Variability of classification rules: criterion type

Schmidt et al. analyzed studies that met specific requirements with regard to the criteria. For instance, in Pearlman et al. (1980), studies using such criteria as turnover, absenteeism, and tardiness were excluded, leaving only job proficiency and training succes criteria. For studies reporting test validities for several criterion dimensions separately as well as for an overall or summary criterion, only the coefficient for the overall or summary criterion was recorded. For cases with several criterion dimensions but with no overall or summary criterion, the average validity over these dimensions was recorded, and the product of the original sample size and the average number of dimensions was taken as sample size. In other publications (Schmidt et al., 1980; Schmidt et al., 1981a; Schmidt et al., 1981b; Callender & Osburn, 1981) again only validities for overall job performance or overall training success were used, and again partial measures (e.g. ratings on specific dimensions of job performance) were excluded.

When job performance was measured by means of several specific dimensions, the composite (sum or average) of these specific criteria was taken as the measure of overall job performance.

This exclusive use of overall criterion measures by Schmidt et al. has non-trivial consequences:

a) Schmidt and Kaplan (1971) demonstrate that ceilings on validity are lower when either the criterion is homogeneous and the predictor is heterogeneous or vice versa, implying lower observed average validity and validity variance (cf. also Smith, 1976, p. 749). In Schmidt et al.'s data base, predictors are classified far more specifically (e.g. in 10 or more types), than the criteria, which are invariably assumed to be of the same global type.

b) Most measures of job proficiency (see e.g. Schmidt et al., 1981 a) were supervisory ratings. From the literature (e.g. De Wolff, 1970) it is well known that ratings may reflect for a major part the personal feelings of the rater, or in the words of Vernon: "there is a strong tendency to evaluate people along the single dimension of how they affect us, and to assume egocentrically that most of their actions are directed towards helping or hurting us" (Vernon, 1964). Thus, with global, overall criteria the validity coefficients may reflect the relations between predictor tests and sympathy for the ratee. This would again lead to little

variance in the distribution of raw validities, this time due to the global, undifferentiated, character of ratings (cf. Schmidt et al., 1984, p. 416).

c) The possibility of multidimensionality of criteria is neglected. Even stronger, Schmidt et al. deny multidimensionality of criteria: ".... only a measure of overall job performance is needed in validity studies ... the only function of multiple criterion scales is to increase the reliability of the composite (overall) criterion measure. That is, replication of judgments on essentially the same dimension leads to increased reliability ..." (Schmidt et al., 1981a, p. 175). However, studies involving statistical analysis of sets of criterion measures rarely yield a single general factor (Smith, 1976). This author mentions the studies of Ewart et al. (1941), Forehand (1963), Grant (1955), Kirchner (1966), Ronan (1963), Rush (1953), Schultz & Siegel (1964), Seashore et al. (1960), Siegel & Pfeiffer (1965), Stark (1959) and Wiley (1964) as evidence against the arguments that unreliability contributes to the "apparent" complexity of criteria. Published analyses, including those on more reliable criteria, lead to accept the conclusion that job performance of employees is as a rule multidimensional (see for instance: Baehr (1967), Brumback & Vincent (1970), Chalupsky (1962), Dowell & Wexley (1978), Fleishman & Ornstein (1960), Fogli et al. (1971), Hemphill (1959,

1960), James et al. (1973), Peres (1962), Prien (1965), Rush (1953), Seashore et al. (1960), Smith & Kendall (1963), Toops (1944), Tornow & Pinto (1976), Wofford (1970)), even though in some cases disattenuated correlations would approach or exceed unity.

Thus, it seems that we have to consider the fact that job performance as a rule tends to be multidimensional, and see what implications this brings along for validity generalization. Traditionally multidimensionality of performance criteria has been dealt with in two ways (Schmidt & Kaplan, 1971; Guion, 1976): (1) component criteria have been isolated and validities have been defined and calculated for each criterion separately, (2) a composite criterion has been defined, measured either directly by some so-called "global" measure, or indirectly by forming a weighted sum of separately measured component criteria. The first approach does not pose special problems for validity generalization, as long as validities against the component criteria are kept separated. Mixing observed validities related to different types of criteria within the same job would, of course, lead to uninterpretable results. The second approach requires special precautions to ensure that the global measures relate to the same composite criterion construct, or sums of separate criteria are properly weighted. The use of observed validities relating to different criterion-constructs would again lead to meaningless results.

Related to this point is the well-established fact that different jobs may have common job elements or criterion dimensions (see e.g. Guion, 1965; McCormick, 1976). The question arises what these relationships between jobs could mean in the context of validity generalization. Following the MTMC-model validity generalization might be applied to data on different jobs, but only when these jobs contain a common criterion construct η, are sought by applicants from a common population Π, and the same predictor construct ξ applies. This means that one should collect and process job component or job element validities, a method already suggested by Ghiselli in 1959. A recent review on three decades of personnel selection research (Monahan & Muchinsky, 1983) reveals that while several authors (e.g. Guion, 1961; Dunnette, 1963) have recommended the use of component criteria, researchers have generally not yet responded to this recommendation.

Concluding, the rules applied by Schmidt et al. for compiling-classifying validity data are rather variable. In some cases validity data on a single test are brought together, in other cases data on parallel tests, tests sharing a common factor, or tests sharing either closely or remotely related factors. The criteria have a global or composite character in most cases, referring to overall success in specific jobs, job types, true job families, or non-true job families, and sometimes also training programs.

Neglect of conceptual homogeneity

The example of Pearlman et al. (1980) makes clear that Schmidt et al. violate the psychometric assumptions of the STSC validity generalization model, by using non-identical tests and criteria. They proceed from a (unspecified) model for which a certain degree of relatedness of predictors and criteria would be sufficient (like the MTMC-model). But, even if one accepts the latter approach for the moment, it may be questioned whether the hundreds of coefficients classified into one test-job type cell are sufficiently related to be considered as operationalizations of the same theoretical validity. As these data may refer to numerous types of tests and to different jobs in different organizations and may relate to different criterion constructs and applicant populations it is doubtful whether they permit validity generalization, even in a MTMC framework.

Schmidt et al. suggest that this classification problem can be solved empirically by using their 'test of situational specificity'. We leave the discussion of this "test" for the next section.

The preceding sections make clear that in the Schmidt et al. procedure the question of conceptual homogeneity of classes is neglected. In those cases where tests within a given class are not identical, they are not, or only very superficially, evaluated in terms of their relationship to an underlying construct. Criteria are never evaluated in

terms of underlying constructs; they may or may not refer to identical or related performances. In the same vein, samples are generally not evaluated in terms of their representativeness for a given population. This state of affairs is at least remarkable, because from a generalization viewpoint a case in which test and criterion are fixed, and the samply is varying, is quite different from a case in which only the test is fixed, and both the criterion and the sample vary. And so on. Failing to distinguish between such cases, by assuming that 'anything goes', makes it unclear what one is generalizing from and what one is generalizing to.

Concluding: the logic of compiling and processing validity data referring to different criterion constructs and populations should be seriously questioned. Earlier researchers, like Ghiselli, may be excused for having done so, because of the moderate level of sophistication in criterion and selection theory at the time. After the conceptual and methodological contributions to this field from Guion (1965), McCormick (1976), Ronan & Prien (1971), Schmidt & Kaplan (1971), and many others, such a way of working seems no longer adequate. In our view it should be avoided, to make sure that meaningful results may be found. To say this with a simple rule: One should only try to generalize data that are, logically, generalizable. There is no alternative to an adequate prior classification of

observed validity data on conceptual grounds. This does not mean, of course, that the subject of job classification is not open to empirical investigation. Such research requires an appropriate methodology however (see e.g. Arvey et al., 1979; Lissitz et al., 1979).

3.2. Generalizability testing

The test of generalizability, or 'non-situational specificity', focuses on the residual validity distribution. The mean of this distribution is the average observed validity. Its variance, called 'residual variance' (S^2_{res}), is defined as the difference between the observed validity variance (S^2_{obs}) and the variance that can be attributed to artifacts (S^2_{art}) : $S^2_{res} = S^2_{obs} - S^2_{art}$. The artifactual variance is estimated by some procedure that takes into account, among other things, (a) the estimated true validity, (b) the sampling error associated with the average sample size, (c) assumed distributions of reliability and selection ratio's (see appendix A for a description of the procedure).

Recently, the various procedures for estimating the artifactual components of validity variance due to effects of attenuation, restriction of range, and sampling error, were scrutinized in a discussion between Callender & Osburn (1980, 1982), Callender et al. (1982), Hunter et al. (1982) and Schmidt et al. (1982). The discussion focused on three methods of estimating the artifactual variance components

from observed validity data: a 'non-interactive' equation proposed by Schmidt et al. (1979), an 'interactive' equation of Schmidt et al. (1980), and the 'multiplicative independent' equation presented by Callender & Osburn (1980). It appeared from simulation studies that all three methods, although not being exactly correct, were reasonably accurate. This conclusion was confirmed by a study of Burke (1984), in which in addition to these three methods a number of other computational procedures were investigated. Because of these findings, the computational procedure(s) for estimating artificial variance(s) in validity generalization are not discussed in this paper.

Generalizability within the data set is established by evaluating either the residual distribution itself, or a transformation of it, the so-called 'prior' distribution of the true validities. The mean of this latter distribution is taken as an estimate of the true validity; it is the average validity corrected upward for restriction of range and attenuation. Its standard deviation is equal to the S_{res}, multiplied with the same correction factor (cf. appendix A).

With respect to the test procedure several comments are in order.

Incorrectness of conceptual basis

First of all, it seems that the conceptual basis of the test is incorrect. Pearlman et al. (1980), in explaining the

principles of the test, state that, given a certain true population validity, artifacts like sampling error, differences in criterion and test reliabilities and restriction of range, may produce observed validity differences of the same magnitude as are found in a given data set. In such a case, generalizability is clearly present. Next, they reverse this argument, asserting that generalizability is present, whenever the observed validity variance is matched by the variance predicted from artifacts. It can easily be seen that this logic is incorrect: from the fact that generalizability leads to $S^2_{res} = 0$, it may not be concluded that any $S^2_{res} = 0$ indicates generalizability. Several combinations of a single true validity and artifactual distributions may underly a given set of observed validity coefficients, resulting in the same observed validity variance. When the assumption of a single true validity is dropped, even more possibilities exist. As a result of too loose a classification, there may quite well be a mixture of two or more populations involved. In section 4.2 it will be demonstrated that even in that case the observed validity variance may be equal to the variance expected on account of artifacts only.

Moreover, according to Schmidt et al. (1979; p. 267) testing the hypothesis of no situational specificity "is conceptually identical to research aimed at establishing general principles about trait-criterion relationships to be used in theory construction". If the situational specificity

hypothesis is rejected "then it follows that various constructs (...) have invariant population relationships with specified kinds of performances and job behaviors". As we have seen, the focus of the Schmidt et al. procedures is, however, the true validity $\rho_{T_x T_y}$ which is test specific. Hence their actual procedure for testing situational specificity does not match their aim of research as stated above. Logically, in the STSC model, the true validity cannot be generalized beyond the specific predictor and criterion measures since it is a correlation between the true score components of these very measures.

Conceptually, testing situational specificity is possible only when the MTMC model is assumed, i.e. when the validity to be generalized is the theoretical validity $\rho_{\xi\eta\Pi}$.

The hypothesis to be tested then, is that validities have been computed on samples from the same reference population. When the test detects differences, it is to be concluded that this hypothesis is not correct, i.e. that the validities come from different populations. In such a case the validity may be concluded to be situationally specific.

Elasticity of decision rules

According to Schmidt et al. generalizability is not only present when $S^2_{res} = 0$. They view generalizability as a matter of degree, depending on properties of the residual distribution and/or the prior distribution (Schmidt &

Hunter, 1977). In fact, several decision rules have been introduced in subsequent publications. Generalizability has been said to be present when:

I there is no residual variance:
$S^2_{obs} - S^2_{all\ artifacts} = 0$ (Schmidt & Hunter, 1977);

II four artifacts explain at least 75% of the observed variance: $(S^2_{obs} - S^2_{4\ artifacts}) / S^2_{obs} < .25$ (Pearlman et al., 1980);

III the 90% credibility value is larger than zero: 90% CV > 0 (e.g. Pearlman, 1982);

IV the 90% credibility value exceeds some 'minimum useful level' u: 90% CV > u (e.g. Schmidt & Hunter, 1977);

V the 90% credibility value exceeds some 'substantial value' v: 90% CV > v (e.g. Pearlman, 1982).

The simultaneous adoption of these decision rules lends the test procedure a great deal of elasticity, allowing the conclusion of generalizability to be drawn in almost any case.

Openness to unknown error

Apart from this, the procedure lacks an underlying sampling distribution, thus precluding the specification of Type I and Type II errors. Callender & Osburn (1981) have tried to solve this problem by generating sampling distributions with the help of computer simulations. However, they have, just like Schmidt et al., equated the null-hypothesis with the hypothesis of interest, which is

not correct. A researcher who is interested in generalizability, should start from a null-hypothesis that assumes non-generalizability. Only in this case does an α indicate the probability of a wrong conclusion (see Hays, 1973, chapter 9 for a general discussion of this statistical problem).

For this reason, the Schmidt et al. test procedure is open to unknown error. The test may lead to an unjustified decision in favor of generalizability with a chance that depends on the type of decision rules applied by the researcher. Recently, Osburn et al. (1983) have subjected the procedure (decision rule no.II) and their own test to an evaluation, making use of computer simulation again. Their results imply a clear warning: the power (i.e. the chance to detect true validity differences) of both procedures was found to be low for the usual condition of small to moderate true validity differences and sample sizes below 100.

Crudeness and unreliability

By its nature the Schmidt et al. test procedure is quite crude, as it evaluates just one attribute of a set of data. It has been shown to be insensitive to changes within the data set, as substantial numbers of deviating (e.g. zero or negative) validities may be added without changing the conclusions (Schmidt et al., 1981a; Callender & Osburn, 1981).

A related point is that the procedure is uninformative: it does not yield information on outliers or potential subclasses. For this reason the test is practically worthless for refining an initial classification.

Of special interest is the fact that the residual variance statistic S^2_{res} seem to be unreliable. Pearlman et al. (1980; p. 384) acknowledge that while S^2_{obs} is susceptible to sampling error, S^2_{art} as calculated by them is only an approximation of the real artifactual variance, based on assumed distributions of artifacts. For these reasons, the difference $S^2_{obs} - S^2_{art}$ may give a wrong indication of the true S^2_{res}. This lack of statistical reliability is clearly visible in those rather frequent instances where $S^2_{art} > S^2_{obs}$ (in the study by Pearlman et al., 1980: 8 out of 32 for proficiency criteria, 7 out of 24 for training criteria, considering only separate job categories; see table 1).

Table 1. Cases where > 100% of observed variance is explained by 4 artifacts (from Pearlman et al., 1980, tables 5 and 6)

Proficiency criteria

test type	job type[a)]	% variance
general	E	121
verbal	C	129
quantitative	E	260
reasoning	A	103
perceptual	C	122
memory	B	153
motor	C	115
motor	E	120

Training criteria

test type	job type	% variance
general	A	107
verbal	A	184
verbal	C	105
quantitative	A	235
reasoning	A	148
reasoning	C	195
clerical	C	152

a) Job types are A = stenography, typing, filing, and related occupations; B = computing and account-recording occupations; C = production and stock clerks and related occupations; E = public contact and clerical service occupations

Probability of bias

The foregoing remark raises another point of criticism, i.e. the possibility that the test is biased, in the sense that the influences of artifacts are overestimated. Although this cannot be established directly, there are several indications that such a bias is likely to be present. A first indication is that the conclusion of generalizability is drawn in most of the cases reported, irrespective of the nature of the data. A second indication comes from the magnitude of s^2_{art} which seems to be too large in too many cases. Table 1 shows the percentages of observed variance explained by four artifacts in the 15 cases mentioned before. The percentages vary from 103% to 260% for the proficiency criteria, and from 105% to 235% for the training criteria. In a study by Brown (1981) the highest percentage of variance explained is even 271%.

Figure 5 presents an overview of the percentages of observed validity variances predicted by the four artifacts in the Pearlman et al. study. As an example: in table 5 of Pearlman et al. the standard deviation of observed validities in the class of (memory tests, computing and account-recording jobs) is .119, where the predicted artifactual standard deviation is .147. Then the percentage explained is $(.147/.119)^2 * 100 = 153\%$. In such a case of over prediction, Pearlman et al. report "100% of variance accounted for". Figure 5 clearly shows the variance attributed to artifacts to be substantial in many cases.

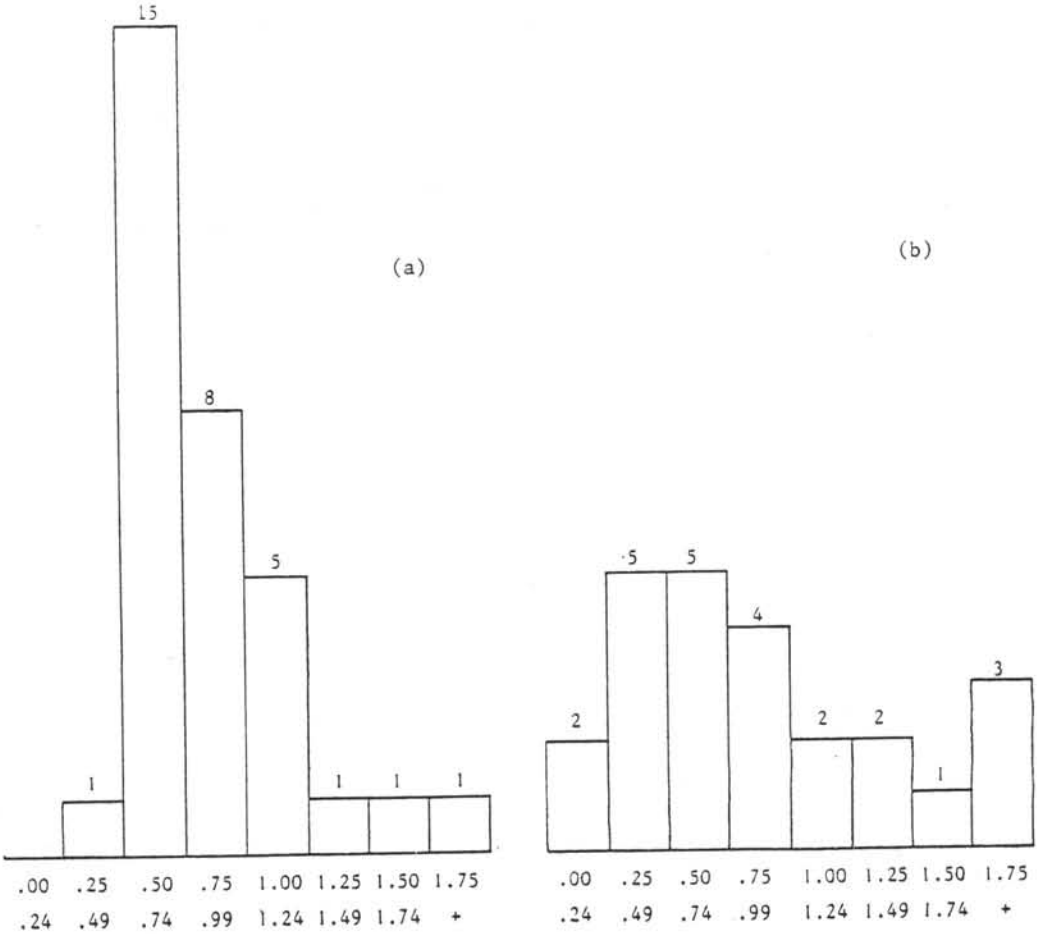

Figure 5. Histogram of percentages of observed validity variance explained by 4 artifacts (from Pearlman et al., 1980, tables 5 and 6).

a : Proficiency criteria (32 cases)
b : Training criteria (24 cases)

One should realize that only four artifacts (restriction of range, criterion attenuation, test attenuation, and sampling error) are involved here. If it were true, as Schmidt et al. have asserted, that other artifacts, such as criterion contamination and deficiency, typographical and data recording errors, etc., account for at least 25% of the observed variance, the percentage of observed variance explained would be over 100 in 50% of all cases reported in the study of Pearlman et al. (1980), as can be inferred from figure 5.

A third indication for bias can be found in a study of Schmidt et al. (1981a, tables 2 and 4), in which validity generalization was applied to several independent (by classification) predictors in a single population of applicants. If it is assumed that the predictors are uncorrelated (as it should be in case of a reliable classification), the multiple correlation with the criterion turns out to lie well above 1 in three of the five job families. It appears that the validities of the predictors have been overestimated in this study.

Thus, it seems that there are serious grounds for the suspection that the test is biased in the direction of generalizability, making S^2_{res} unduly small. A possible explanation may be found in the growing body of evidence that the corrections for attenuation and for restriction of range cannot be advocated in all situations. For instance, Lee et al. (1982) found that by employing these corrections

consecutively, overcorrection may be easily obtained. Apart from this, the robustness of the corrected estimator $r_{T_x T_y P_j}$ for violations of classical test theory assumptions is low (Lumsden, 1976; Winne & Belfroy, 1982). In case of a low reliability, the coefficient can be quite biased, and the standard error will be high (Bobko & Rieck, 1980, p. 395).

In the same way, it has been shown that the correction for restriction of range is not always appropriate. Specific factors affecting this correction are for instance the presence of a low population correlation (smaller than .30; Overbeek, 1974; Greener & Osburn, 1979), violation of the linearity assumption (Greener & Osbur, 1979), the use of unknown, implicit selection rules (Roe & Elshout, 1972; Linn et al.,1981; Gross & Perry, 1983), and the use of a variable cut-off score (Olson & Becker, 1983; Roe, 1983; p. 472-476).

A general conclusion from the preceding discussion is that application of the attenuation and range restriction corrections should be guided by the nature of the predictive validity at issue, i.e. by specific aspects of the validity study. It seems likely, for instance, that the actual selector almost never will coincide with the predictor variable; in such a case, a careful inspection of the selection procedure should be conducted. Roe (1979) proposed to reconstruct the selector by means of a multiple regression of the dichotomous selection criterion on a number of conceivable predictors of selection (e.g. age,

sex, school results). If indeed selection is to a high degree situationally specific, the evaluation of the actual selection procedure necessarly will have to be also situationally specific.

Questionable applicability to heterogeneous data

A final remark concerns the applicability of the test to heterogeneous data, such as may result from some of Schmidt et al.'s loose classification rules. If the test and the criterion to which the validities refer are both fixed, a test on numerical differences of validities may, in principle, allow conclusions on the probability that the samples involved come from a common population. In such a case the test results may help to refine the initial classification. With differing tests, criteria and populations, it is unclear to what state of affairs the result of a statistical test refers.

Schmidt et al.'s test procedure is based on the assumption that differences between observed validities in conceptual referents, e.g. the criterion constructs and populations involved, will be reflected in numerical differences, as only such numerical differences are taken into account. However, it is easy to see that this assumption underlying the test is wrong. Equal or slightly different numerical values are neither a necesary, nor a sufficient condition for the validities being equivalent in the sense that predictor and criterion referents are equal.

The fact that two different tests have equal validities against two different criteria contains little information about the relation between the criterion constructs. Also, at least within a wide range, the same test may have equal validities against criteria having no conceptual or even empirical relationship. Only at very high levels of validity would equal values have implications for empirical relations among criterion measures.

Thus, logical reasons preclude to draw conclusions on the similarity of different predictors or criteria, or on the exchangeability of the validities involved. This implies that generalizability testing without strict classification rules is essentially meaningless.

Concluding: the Schmidt et al. testing procedure seems to violate some methodological principles. Its conceptual basis is doubtful, the decision rule is subjective and it varies over studies, the procedure seems to favor the null hypothesis of no situational specificity, and the power is unacceptably low, type I and type II errors are unknown, the alternative hypothesis is much too diffuse to yield useful practical information, residual variance seems consistently underestimated, robustness issues with respect to correction for attenuation and restriction of range are ignored, and finally, its loose data base precludes that any useful information can be obtained with respect to classification rules.

3.3. Generalization

The third component of the Schmidt-Hunter validity generalization method is a procedure for making inferences on the true population validity. The mean observed validity $\bar{r}$ is corrected for restriction of range and attenuation, producing an estimated true validity: $\hat{\rho}$ (in fact $\hat{\rho}_{xTy\Pi}$ or $\hat{\rho}_{TxTy\Pi}$). The same correction is applied to S_{res}, producing an estimate of the standard deviation of the true validity: $\hat{\sigma}_{\rho}$. Schmidt et al. conceive of a normal 'prior distribution' with $\hat{\rho}$ as mean, and $\hat{\sigma}_{\rho}$ as standard deviation. Both the mean and the lower bound 90% credibility value (90% CV) of the true validity are reported as generalization outcomes.

Below, a number of critical remarks with respect to this procedure are listed.

Incompleteness

A first point of criticism, already mentioned in section 2.2, is that Schmidt et al.'s generalization procedure is incomplete, as it ends with statements on the true validity ρ to be expected on future occasions. This ρ does not relate to the use of a specific test for predicting a given criterion in a new sample of applicants, but rather to the imaginary use of this test (or its perfectly reliable counterpart) for the prediction of a perfectly reliable criterion in a sample of infinite size. This means that the validity estimate obtained should be considered as a

parameter for a prediction model that lies at the level of theoretical constructs. If a model is desired at the level of operational measures, as is usually the case, the generalization procedure should be supplemented. The inductive phase in which $\bar{r}$ is transformed into $\hat{\rho}$, should be followed by a deductive phase in which an estimate is made of a future observed validity, taking into account the actual criterion reliability r_{yy}, the test reliability r_{xx}, and the future sample size (see figure 4). To be sure, restriction of range should be left out, as the parameters needed are those to be applied in a model for all applicants. Of course, the elaboration of the generalization procedure would lead to lower levels of validity and wider credibility intervals, with for some cases trivial results.

Unadaptedness to generalizability differences

A peculiarity of the Schmidt-Hunter et al.'s method is that it is not adapted to differences in generalizability within data sets (cf. Algera et al., 1984). Schmidt et al. seem to distinguish between three cases:

Case I : there is generalizability because the hypothesis of no situational specificity can be accepted;

Case II : there is generalizability although the hypothesis of situational specificity cannot definitely be rejected;

Case III : there is generalizability although the hypothesis of no situational specificity can definitely be

rejected.

In all cases the generalization procedure is essentially the same. The only difference is, that more weight is assigned to the 90% credibility value (instead of $\hat{\rho}$) in the cases II and III. However, it would only seem logical not to generalize in these latter cases, and instead reclassify the data. Another defendable option would be to choose a generalization procedure that deals with two or more true validities (cf. section 5). Using the same procedure can only lead to inaccurate or even false results.

Inappropriateness of psychometric model

It seems that the generalization procedure is based on a very restrictive, and in fact inappropriate psychometric model. This model, which has been derived from classical test theory, has been summarized by Callender & Osburn (1980) as $r = \rho.a.c + e$, in which a is the correction factor for criterion attenuation and c the one for restriction of range, while e is a term for sampling error. Following our notation the model can be rewritten more completely as:

$$r_{xyp_j} = \rho_{xT_y\Pi} \cdot a_{yP_j} \cdot c_{P_j} + e_{P_j} \qquad (9)$$

Equation (9) shows that when different samples are drawn from a population with a given true validity, different r's will be observed. The true validity $\rho_{xT_y\Pi}$ is a constant for each test-criterion-sample combination, while all other

factors may vary over samples. Although this STSC-model leaves no room for more than one ρ at the same time (see figure 1), Schmidt et al. act as if it would apply for several tests, criteria, and populations. This creates insolvable problems of interpretation. E.g. what is exactly the meaning of a $\hat{\rho} = .60$ for a collection of 3 arithmetic reasoning tests, 2 tests for computational speeld, and 1 test for numerical estimation? And what conclusion should be drawn with respect to the use of a computational test, given this datum?

As will be shown in section 5, what would rather be required, is a MTMC-model that explicitly considers the relationships between different tests and underlying constructs, as well as between criterion measures and constructs.

Probability of bias

The remark on the probability of bias, made in section 2.2, in the corrections for restriction of range and attenuation applies here as well. To repeat, there are reasons to assume that there is a bias in the chain of corrections leading to an overestimation of true validity. In this case, the bias will affect the general level of true validities, or of minimum useful true validities, which, in turn, affects the implications for practical selection of validity generalization in the sense of Schmidt et al.

Non-Bayesian character

A Bayesian approach to validity generalization would proceed as follows: in the inductive phase, the parameter of interest is $\rho_{\xi\eta\Pi}$, conceived as a random variable. Prior information relevant to this variable is available in the form of distributions for criterion and predictor reliabilities and the selection ratio, like those presented by Schmidt et al. Initially, a prior probability distribution with respect to $\rho_{\xi\eta\Pi}$ is assumed. The observed raw validities constitute the data. A sampling model is assumed to link these validities to $\rho_{\xi\eta\Pi}$. Next, the prior information and the data are combined to produce the posterior information, i.e. to derive the posterior density for $\rho_{\xi\eta\Pi}$ given the data. From this distribution point estimates and credibility intervals may be derived.

In the deductive phase the attention is focused to a future $r_{x_h y_i P_j}$, which is also considered a random variable. Given the posterior density for $\rho_{\xi\eta\Pi}$ and assuming a sampling model which links $r_{x_h y_i P_j}$ to $\rho_{\xi\eta\Pi}$, a predictive density for $r_{x_h y_i P_j}$ can be derived. The density contains all the information with respect to $r_{x_h y_i P_j}$, and hence can be used to produce point and interval predictions.

Viewing this, it is obvious that Schmidt et al.'s approach, which is labeled 'Bayesian', rests in fact on ideas from classical statistics. In a Bayesian approach prior information on a probabilistic hypothesis is combined with relevant data in order to produce posterior information

on the hypothesis (inductive phase), on the basis of which predictions about future observations may be generated (deductive phase). Essentially, Schmidt et al. have only offered a procedure for arriving at an 'empirical' prior (Schmidt & Hunter, 1977). There is no transition from prior to posterior information. Also the element of setting up a predictive distribution for future observations, based on the posterior (Vijn, 1983), is lacking.

As will be shown in section 4, the neglect of Bayesian principles in predicting a future validity affects the level of the predicted validity in a non-trivial way.

Concluding: the Schmidt et al. generalization procedure has a number of shortcomings: it is incomplete as it ends with an estimate of a true validity, omitting the phase of deducting from this ideal validity the validity to be expected upon practical application in realistic circumstances, it does not take different levels of situational specificity into account, it is based on a restricted psychometric model but used in a much more wider sense, it again is affected by non-trivial bias, which in this case may result in too large a generalized true validity, and finally it is essentially non-Bayesian in character. A Bayesian remodeling of Schmidt et al.'s generalization procedure will be set up, and studied in the next section.

A general conclusion of section 3 is that there is insufficient reason to feel confident with Schmidt et al.'s validity generalization, viewing the inaccurate classification procedure, the conceptually and technically deficient test of situational specificity, and the inappropriate generalization model. A summary of the various points of criticism presented in this section with respect to the three components of validity generalization in the sense of Schmidt et al., can be found in table 2.

Table 2. Evaluation overview

---

1) *compilation-classification*
   - variability of classification rules: test type, job type, criterion type
   - neglect of conceptual homogeneity

2) *generalizability testing*
   - incorrectness of conceptual basis
   - elasticity of decision rules
   - openness to unknown error
   - crudeness & unreliability
   - probability of bias
   - questionable applicability to heterogeneous data

3) *generalization*
   - incompleteness
   - unadaptedness to generalizability differences
   - inappropriateness of psychometric model
   - probability of bias
   - non-Bayesian character

---

4. Remodeling Schmidt et al.'s generalization procedure

In this section we will demonstrate how the generalization component from Schmidt et al.'s method may be remodeled by applying the framework of Bayesian analysis to their psychometric model. The resulting procedure is described in section 4.1. In section 4.2 then, the question is raised whether this procedure is sufficiently robust to justify the violations of the assumptions on test type and criterion type that seem frequently present in Schmidt et al.'s studies. This issue will be clarified by simulated data.

4.1. Bayesian remodeling

The general scheme of Bayesian analysis presented above can be applied to the STSC model as follows. The central parameter is the population true validity $\rho_{T_x T_y \Pi}$; it will be denoted as ρ from here on. From populations Π a 'selected' population π may be obtained, selecting on basis of X. Within this subpopulation, a 'raw' validity $\rho_{xy\pi}$ may be defined, which is consequently related to ρ by a function $g(\rho, \rho_{xx}, \rho_{yy}, \kappa)$:

$$\rho_{xy\pi} = \frac{\sqrt{\rho_{xx}\rho_{yy}}\,\kappa}{\sqrt{\rho_{xx}\rho_{yy}\,(\kappa^2-1)\rho^2 + 1}}\;\rho \qquad (10)$$

representing the combined effects of attenuation and restriction of range on ρ. Parameters ρ_{xx} and ρ_{yy} are the reliabilities of the specific predictor X and criterion Y respectively (indices h and i are omitted since there is only one predictor and one criterion instrument now), whereas κ indicates the degree of range restriction: κ is equal to the ratio of the restricted standard deviation of X to the unrestricted standard deviation of X.

In actual 'selected' samples p_j from the specified π, a correlation r_{xyp_j}, further denoted as r_j, instead of $\rho_{xy\pi}$ will be observed. Differences between $\rho_{xy\pi}$ and the observed validities r_j are assumed to be the result of sampling error only (Callender & Osburn, 1980). According to Callender and Osburn (o.c., p. 548) "the objective of the validity generalization analysis is to determine the mean and variance of ρ ". Note that this restricts the analysis to the induction of a true validity $\rho_{T_x T_y \Pi}$.

At the prior stage, the available information on the parameters ρ_{xx}, ρ_{yy}, κ is represented by prior densities. Schmidt et al. assume that normal densities apply (cf. e.g. Pearlman et al., 1980, p. 375f).

At the sampling stage the observed validities r_j are assumed independent realizations of a random variable R_j, the distribution of which, according to Callender & Osburn (1980), may be supposed normal with mean $\rho_{xy\pi}$ and variance v_j, where v_j may be approximated (Pearlman et al., 1980) by

$$\frac{(1 - r_j^2)^2}{n_j}. \qquad (11)$$

(Schmidt et al's procedure could be improved here by applying Fisher's Z transformation on the r_j in (11)). Since $\rho_{xy\pi}$ is the function g (.) of the STSC parameters, the final sampling distribution can be denoted as $p(R_j \mid \rho, \rho_{xx}, \rho_{yy}, \kappa)$. It could be approximated in frequency from when samples with constant size n_j would be repeatedly drawn from the reference population.

At the posterior stage, the posterior density of ρ given the data is derived using Bayes' Theorem. In appendix B it is shown that this posterior density may be satisfactorily approximated by a normal density

$$\rho \mid \text{data} \sim N(\tilde{\rho}, \tilde{\sigma}_\rho^2). \qquad (12)$$

Estimation procedures for the posterior mean $\tilde{\rho}$ and variance $\tilde{\sigma}_\rho^2$ are described in appendix B as well. An important result is that $\tilde{\rho}$ can be estimated by solving the implicit equation

$$g(\tilde{\rho}, \rho_{xx}, \rho_{yy}, \kappa) = \sum_{j=1}^{K} w_j r_j, \qquad (13)$$

for $\tilde{\rho}$, where the weights w_j are defined as

$$w_j = \frac{\frac{1}{v_j}}{\sum_{j=1}^{K} \frac{1}{v_j}} . \qquad (14)$$

and the parameters ρ_{xx}, ρ_{yy}, κ should be replaced by appropriate estimates.

The procedure actually given by Schmidt et al. is somewhat different. Instead of (12), they use the density

$$N(\hat{\rho}, \hat{\sigma}_{\rho}^{2}). \qquad (15)$$

Details for the estimation of the mean $\hat{\rho}$ and the variance $\hat{\sigma}_{\rho}^{2}$ can be found in Schmidt et al. (1980) (see also appendix A). For the present exposition, it is important to note that the mean true validity $\hat{\rho}$ is estimated by solving the equation

$$g(\hat{\rho}, \bar{\rho}_{xx}, \bar{\rho}_{yy}, \bar{\kappa}) = \sum_{j=1}^{K} w_j r_j, \qquad (16)$$

with weights

$$w_j = \frac{n_j}{\sum_{j=1}^{K} n_j} . \qquad (17)$$

So in the actual Schmidt et al. method, the weights w_j are defined in terms of the sample sizes n_j, in stead of in the "precision" v_j as should be in a full Bayesian approach.

The weighting function (17) is non-optimal in terms of minimum mean-square error properties (Efron & Morris, 1977; Lindley & Smith, 1972). Furthermore, in (16) the uncertainty as embodied in the prior densities for ρ_{xx}, ρ_{yy}, and κ has no influence on the point estimate of ρ, as in (18) prior means $\bar{\rho}_{xx}$, $\bar{\rho}_{yy}$, and κ are inserted irrespective of the variances of these densities. Thus, although our remodeling of the Schmidt et al. procedure deviates somewhat from the actual method at this point, it appears to be a refinement of the original formulation. Therefore, we will continue to apply (12) in stead of (15) in our Bayesian remodeling of Schmidt et al.'s procedure.

In the deductive phase, the variable of interest is a future $r_{x_h y_i P_j}$. Schmidt et al. disregard criterion unreliability, so the validity of interest becomes $r_{x_h T_{y_i} P_j}$ denoted as r_f. Given the posterior density for ρ in (12), and denoting the future sample size as n_f and the future test reliability as r_{xx}, the predictive density for r_f appears to be (cf. appendix B)

$$N(\sqrt{r_{xx}}\,\tilde{\rho},\ r_{xx}\,\tilde{\sigma}_{\rho}^{2} + \frac{(1 - r_{xx}\,\tilde{\rho}^{2})^{2}}{n_f}) \ . \qquad (18)$$

This density contains, in probability form, the information available with respect to the value of the validity to be observed in the future study.

Formula (18) represents a proper Bayesian remodeling of the Schmidt et al. procedure. Actually, Schmidt et al. present in stead of (18) the density

$$N(\sqrt{r_{xx}}\,\hat{\rho},\ r_{xx}\,\hat{\sigma}^2_{\rho}), \tag{19}$$

which is (15) corrected for predictor unreliability. Obviously, the Bayesian Schmidt et al. procedure (18) and the actual Schmidt et al. procedure (19) differ with respect to estimation of the mean and variance of the future validity. Suppose $\tilde{\rho} = \hat{\rho}$ and $\tilde{\sigma}^2_{\rho} = \hat{\sigma}^2_{\rho}$. Then the Schmidt et al. variance of R_f will be smaller than the variance estimated by the proper Bayesian procedure. Comparison of (19) to (18) shows that the term $(1 - r_{xx}\,\tilde{\rho}^2)^2 / n_f$ is ignored in the former procedure, which is equivalent to assuming that n_f is infinite. So our remodeling takes into account the predictive uncertainty due to a finite sample size. As a consequence, interval estimates of a future validity are less narrow and 90% lower bound credibility levels of validity are lower.

For example, consider the Verbal Ability (A) job category in Pearlman et al.'s table 7 (1980, p. 388 f) : $\tilde{\rho} = \hat{\rho} = .39$, and $\tilde{\sigma}_{\rho} = \hat{\sigma}_{\rho} = .23$. Suppose r_{xx} is equal .6, and n_f equal 60. Pearlman et al. report a minimum <u>true</u> validity of .10. Taking predictor unreliability into account .10 drops to .074 (attenuation), and taking the size of the future sample into account .074 drops to .021. Though

.10 may have some relevance, .021 seems practically useless for selection purposes.

4.2. Robustness of STSC in a Bayesian framework

As stated above, Schmidt et al., although implying the STSC-model by their procedure, have consistently used the STSC-model as if it were a MTMC-model. Hence, the question should be considered to what extent and in which way violation of the STSC-model assumptions affects the results of validity generalization in an STSC-framework.

Assume that the STSC model is violated in the sense that in stead of one true ρ that is generalizable over the studies, there a two true validities ρ_1 and ρ_2 "underlying" the K observed validities. In this situation a Bayesian procedure can be set up analogous to the one described above (cf. Jansen et al., 1984, for details). The inductive phase of deriving estimates of ρ_1 and ρ_2 proceeds completely analogous to (12) - (14). However, in order to formulate a predictive density for r_f in the deductive phase, the assumption will be made that with probability c the future study is a sample from the ρ_1-population, and with probability 1-c a sample from the ρ_2-population. Under this assumption, the predictive density for r_f can be approximated by a mixture of two normal densities of the form (18), but with $\tilde{\rho}_1$ or $\tilde{\rho}_2$ substituded for $\tilde{\rho}$: (cf. appendix C):

$$\rho(r_f \text{ all previous studies}) = c * p\ (r_f \text{ type } \rho_1\text{-studies}) + (1-c) * p\ (r_f \text{ type } \rho_2\text{- studies}). \quad (20)$$

Procedures for estimating the mean and variance of r_f for such a mixture of two normals are described in Everitt & Hand (1981).

Using (20), a minimum level of validity r^* with confidence α can be obtained from solving

$$Pr(r_f > r^*) = \alpha. \quad (21)$$

This is the minimum level of validity to be expected with $(1-\alpha)$% confidence in the future study. Below, this Bayesian treatment of the deductive phase of validity generalization in case of a mild form of situational specificity is compared to the correponding STSC-procedure, als described above. The comparison is quantitative to gain some insight into the actual effects of using the STSC-model when it is, in fact, violated. The Bayesian remodeled Schmidt et al. STSC-procedure will be used, including the refinement on that method presented in the previous section. The approach based on (20) will be denoted 'MTMC' in the sequel.

In the STSC approach, the predictive density is given by (18), which is unimodal. This is understandable since in STSC one operates as if situational specificity had been rejected, thus a single underlying ρ is conceived. In the MTMC approach, there are 2 true validities, ρ_1 and ρ_2, which implies that the predictive density (20) very probably

will be bimodal. Of course, with a bimodal distribution of observed validities the use of STSC-analysis is contra-indicated. Consequently, to impart some practical importance on the exercise in this section, it will be assumed that the density (20) is unimodal (which makes the violation of STSC rather mild, at first view that is).

Assume that the predictive density (20) is unimodal with mean $\tilde{r}$ and variance σ^2:

$$R_f \mid \text{data} \sim \text{unimodal } (\tilde{r}, \sigma^2). \tag{22}$$

Sufficient conditions for unimodality of the mixture (20) can be found in Everitt and Hand (1981). Furthermore, we assume that the Bayesian predictive density (22) is matched as closely as possible to the predictive density (18) that follows from the STSC approach. This can be achieved by choosing $\tilde{r}$ and σ^2 equal to the mean and variance of (18). To enhance comparison to some actual Schmidt et al. data, it will be assumed that in (21) $\tilde{\rho} = \hat{\rho}$ and $\tilde{\sigma}^2_\rho = \hat{\sigma}^2_\rho$.

The the entire matching process can be explained by considering an example. In table 7 of Pearlman et al. (1980, p. 388) we find that for the General Mental Ability and Type A - test/job category $\hat{\rho} = .50$, $\hat{\sigma}^2_\rho = .24$, and the 90% point is .19. Taking into account $r_{xx} = .80$, the attenuated predictive density (19) becomes N(.447, .04608) leading to a 90% point of .17. Taking subsequently into account a future sample size of $n_f = 50$, the predictive density (18) becomes

N(.447, .0589), leading to a 90% point of .136.

Switching to the MTMC approach we have to generate a unimodal mixture of two normals $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with mean $\tilde{r} = .447$ and variance $\sigma^2 = .0589$:

$$cN(\mu_1, \sigma_1^2) + (1-c)\, N(\mu_2, \sigma_2^2). \qquad (23)$$

To get an impression of the power of the MTMC model in estimating a 90% minimum useful level of validity, a series of unimodal mixtures (23) has been generated. The procedure was as follows:

1. Select a μ_1 from the range (0, .42) and μ_2 from (.44, .80), and compute c such that the overall mean is equal to .447.
2. Select σ_1^2 and σ_2^2 such that the overall variance is equal to .0589.
3. Check the sufficient unimodality conditions (Everitt & Hand, 1981).
4. Compute a 90% minimum level for $N(\mu_1, \sigma_1^2)$, denoted as r_I.
5. Compute a 90% minimum level for $N(\mu_2, \sigma_2^2)$, denoted as r_{II}.
6. Compute the weighted mean $r_{mean} = cr_I + (1-c)r_{II}$.
7. Check whether the mixture shows sufficient resemblance to a normal distribution.

Some results of the simulation have been listed in table 3.

Table 3. Means and 90% points in a unimodal mixture of two normals (MTMC), compared to the corresponding STSC results.

MTMC predictive						STSC predictive		
c	μ_1	μ_2	r_I	r_{II}	r_{mean}	minimum true validity (Schmidt et al.)	minimum, corrected for predictor unreliability	minimum, corrected for pred. unrel. and for future sample size
.006	.00	.45	-.353	.114	.111	.19	.17	.136
.180	.12	.52	-.160	.267	.189	.19	.17	.136
.491	.32	.58	.024	.307	.163	.19	.17	.136
.928	.42	.80	.109	.519	.138	.19	.17	.136

From table 3 the following can be inferred. When e.g. μ_1 = .12 and μ_2 = .52 then with probability c = .18 the predicted 90% point is - .16, and with probability .82 a 90% point of .267 will be obtained. When the future validity study happens to be in reference population I, a negative 90% point may be expected. If c = .491 a 90% point of .024 is predicted for population I, implying no validity at all, and with probability, .509 the 90% point equals .31. implying substantial validity.

Some general conclusions from table 3 are:

First the mean lower bound r_{mean} may differ substantially from the STSC lower bound, regardless whether the latter is

given as the true validity (.19) as Schmidt et al. do, or as the more appropriate minimum level corrected for both predictor unreliability and size of the future sample (.136). Note that table 3 gives only a sample of the simulation results.

Second, the lower bound to be expected over both reference populations I and II may, under certain circumstances, be approximately equal to the proper STSC lower bound. However, in such cases (e.g. c = .491 in table 2) a substantial increase in minimum validity can be obtained by correctly identifying the type of the validation study. For instance, identifying type II studies when c = .491, heightens the minimum validity from the average .16 to .31. Thus, the simulation study confirms that when the assumption of situational non-specificity is, even mildly, violated, a more thorough classification procedure can yield a substantial gain in expected validity.

And, of course, the reverse holds: Knowing that a study belongs to type I may yield the conclusion that no validity is to be expected in a future study (e.g. situation c = .491 in table 3). Nevertheless, in such a case, the STSC lower bounds can be deceptively large.

Concluding: Remodeling the Schmidt et al. generalization component according to a Bayesian framework implied the addition to the procedure of a deductive phase in which a "real-life" validity is predicted. The analyses in section

4.1 showed that non-trivial differences are obtained between this validity and the true validity estimated by Schmidt et al., which holds the more when the generalization procedure is refined by taking uncertainty with respect to future sample size into account. In section 4.2. this remodeled and refined STSC-generalization procedure appreared to be insensitive to a MTMC data structure. When the data were properly analyzed (by a unimodal-mixture-of-two-normals-Bayesian-MTMC-model) generalizability appeared to vary between zero or even negative validity and substantial validity for future studies.

5. Validity generalization revisited

The discussions in sections 2 and 3 made it clear that the three components of Schmidt et al.'s procedure for validity generalization, viz. classifying validity data, testing the hypotheses of situational specificity, and generalizing validities, are open to serious criticism. On itself such a finding may not be seen as compelling with regard to the procedure's usefulness. Some might consider the procedure as sufficiently "robust" to overcome the multitude of problems listed in sections 2 and 3 in practice. However, the investigation of section 4 showed that the Schmidt et al. method, even when remodeled in a proper Bayesian sense and even when refined at a non-trivial point, definitely does not posses such a robustness. This prevents the application of the rather strict STSC-model of

validity generalization on data that violate (even mildly) the assumptions of such a model.

The foregoing implies that the Schmidt et al. procedure for validity generalization, both for conceptual, theoretical and for practical reasons, leaves room for improvement. In this section we will give suggestions for improving the three components of validity generalization.

5.1. Compilation-classification

It was concluded in section 3.1 that viewing the conceptual and technical deficiences of the "test" for situational specificity, which was proposed as a check on the data compilation, there is no alternative for a thorough a priori classification of validity data. Such a set of compilation rules many well result in posing more constraints on validities classified into one "test-job-sample type" cell than was the case with Schmidt et al.'s procedure.

With the STSC-model, which only allows generalizations to future observations of the same predictor-criterion combination, the proper rule would be to limit generalization to validities referring to the same predictor measure and the same criterion measure observed in different samples from the same population. However, taking the MTMC-model als the base of validity generalization, the proper rule would be that the measures X_h and Y_i are adequate operationalizations of the latent predictor construct ξ and

criterion construct η. This leaves room for compiling validites on varying predictor-criterion combinations into one test/job category, provided that both predictor variables and criterion variables have a sufficient loading (we would suggest $\geq .70$, as in that case construct and variable share at least 50% variance) on their respective constructs ξ and η. We would furthermore suggest the requirement that the constructs ξ and η are each measured by at least three independently constructed but statistically congruent operational measures X_h or Y_i.

As with the STSC-model, the MTMC-conceptualization of validity generalization requires that the population referent Π is the same for all validities classified into an homogeneous cell. The equivalency of validities with respect to such general and often reported population referents, as e.g. age or sex of subjects, is easily verified, but for such referents as e.g. educational background this may be not so evident. In such a case, one may (and often will) choose to ignore the referent.

Still, there may remain cases in which it is unclear whether a referent that differs between validity studies (e.g. sex) influences the validities in such a way that they cannot be classified as homogeneous. Essentially, answering such a question implies checking the situational specificity of the pertinent predictor-criterion relationship for the population referent at issue. Some means for testing this hypothesis of situational specificity are suggested in the

next section.

5.2. Generalizability testing

The hypothesis to be tested is that validities with different population referents are equal, i.e. can be considered as samples from the same population. The most direct way to test this hypothesis would be to investigate whether the validities that are obtained after correcting the observed validities for artifacts, are equal to each other. This would lead to a statistical test of

$$H_0 : \rho_1 = \ldots\ldots = \rho_K,$$

with as an alternative that the ρ's are unequal. Compared to the "test" of Schmidt et al. such a test would be expected to be sensitive to departures from H_0, and to have a high power of detecting true differences. The test would be approached by some homogeneity of correlations test (Hays, 1973; Viana, 1980; Kraemer, 1979). Existing tests, however, only deal with sampling error, and the effects on the standard errors of the correlations due to unreliability of predictor and criterion, and of range restriction have have not yet been fully worked out. It is known that the sampling variance of a correlation corrected either for unreliability (attenuation) or range restriction is larger than the sampling variance of the uncorrected correlation (Bobko & Rieck, 1980; Roe, 1983). Empirical approximations to the

standard error of a correlation in case of restriction of range are given by Forsyth (1971). When there is evidence that validities are near zero, the standard error of the restricted correlation may be approximated by the standard error of the corresponding unrestricted correlation (Elshout et al., 1979). Forsyth and Feldt (1969) derived empirically standard errors for a correlation corrected for attenuation.

When the tests points to significant differences in validities, a subsequent investigation might reveal specific referents on which the validities differ systematically (e.g. educational differences between samples) and that act as moderators therefore of the predictor-criterion relationship. This may help to refine the classification scheme, but also the prediction system, as was illustrated by the simulation study in section 4.2.

5.3. Generalization

In stead of the strict STSC-model, one may chose to take the MTMC-model, which allows looser classification and wider generalization, as a basis for validity generalization.

Analogously to the procedure for the STSC-model as described in section 4.1., a two stage full Bayesian validity generalization method can be set op for the MTMC-model. In the latter model, the central parameter of interest is the theoretical validity $\rho_{\xi\eta\Pi}$ (see figure 2), abbreviated as ρ. Besides it, a "raw" population validity $\rho_{x_h y_i \Pi}$ may be defined: the validity that is

obtained when, instead of the constructs ξ and η, a pair of operational measures X_h and Y_i is used. As there are many of such pairs, we write $\rho_{x_h y_i \Pi}$.

Using the model equations (4) and (5) from section 2, the relationship between $\rho_{x_h y_i \Pi}$ and ρ is:

$$\rho_{x_h y_i \Pi} = \beta_h \gamma_i \rho. \qquad (24)$$

β_h and γ_i are standardized coefficients indicating the validities of test X_h and criterion Y_i against their respective constructs ξ and η .

It should be noted that there is an analogy between (24) and the attenuation correction formula from classical test theory, β_h corresponding to $\sqrt{\rho_{xx}}$ and γ_i to $\sqrt{\rho_{yy}}$. Drasgow and Miller (1982) advocate the use of β_h and γ_i as measures of instrument validities, because the traditional correlations between test score and true score (i.e. $\rho_{xT_x} = \sqrt{\rho_{xx}}$ and $\rho_{yT_y} = \sqrt{\rho_{yy}}$) lack a substantive interpretation.

The derivation of (24) is staightforward and need not be presented here. It may also be found by applying path analysis in figure 2: the correlation between X_h and Y_i is equal to the product of all coefficients belonging to the path that connects X_h and Y_i.

As observed validities tend to be obtained under conditions of range restriction, the range restricted counterpart of $\rho_{x_h y_i \Pi}$ i.e. $\rho_{x_h y_i \pi}$ should be considered:

$$\rho_{x_h y_i \pi} = \frac{\kappa_j \rho_{x_h y_i \Pi}}{\sqrt{(\kappa_j^2 - 1)\ \rho^2_{x_h y_i \Pi} + 1}} \quad . \tag{25}$$

κ_j represents the degree of range restriction in validity study j.

From (24) and (25), a "combined" formula can be obtained, which can be written for short as the function $g^*(\rho, \beta_h, \gamma_i, \kappa_j)$ of the MTMC-model parameters. $\rho_{x_h y_i \pi}$ is the expectation of $r_{x_h y_i p_j}$, i.e. the validity that would be expected over repeated sampling with a fixed sample size n_j and fixed instruments X_h and Y_i.

A Bayesian procedure may be defined analogous to the one described in section 4.1. Thus, prior densities for ρ, β_h, γ_i and κ_j may be specified, and a posterior density for ρ as well as a predictive density for the future validity r_f may be derived. In case of the MTMC-model, we have to make assumptions about the functional form of K prior densities $p(\beta_h)$, one for each validity study; the same holds for $p(\gamma_i)$ and $p(\kappa_j)$. Or in other words: for each given predictor instrument X_h, we have, in distributional form, a "belief" about the value of the corresponding instrument validity β_h.

However, on account of the classification rules proposed in section 5.1 which guarantee a certain homogeneity within test type/job category cells (ξ, η), it may be safely

assumed that all these K "beliefs" $p(\beta_h)$ have the same functional form. Technically, this can be realized by assuming the stochasts β_h to be "exchangeable": they are assumed realizations from one random variable β stemming from a "super"-population. Exchangeability is implied by random sampling, but not vice versa. For models based on exchangeability of parameters, see Lindley and Smith (1973), Rubin (1980, 1981), Vijn (1980).

Thus:

$$\beta_h \sim N(\overline{\beta}, \sigma_\beta^2) \qquad (h = 1, \ldots, L)$$

$$(X_h, Y_i \text{ pertaining to } \xi, \eta),$$

$$\gamma_i \sim N(\overline{\gamma}, \sigma_\gamma^2) \qquad (i = 1, \ldots, M) \qquad (26)$$

$$\kappa_j \sim N(\overline{\kappa}, \sigma_\kappa^2) \qquad (j = 1, \ldots, K)$$

The prior means and variances are assumed to be known either from previous studies or from subjective considerations. Because of the restrictive classification rules, $\overline{\beta}$ and $\overline{\gamma}$ are high, .70 at least. Furthermore, since we are 95% sure that β_h lies between .60 and .90, $4 \times \sigma_\beta \approx .30$ so that $\sigma_\beta^2 \approx .0056$. The same reasoning can be applied to $\overline{\gamma}$ and σ_γ^2, and to $\overline{\kappa}$ and σ_κ^2.

Assuming a non-informative prior for ρ, the joint posterior density is

$$p(\rho, \beta_1, \ldots, \beta_L, \gamma_1, \ldots, \gamma_M, \kappa_1, \ldots, \kappa_K \mid \text{data}) \propto$$

$$\prod_j p(r_j \mid \rho, \beta_h, \gamma_i, \kappa_j)\, p(\beta_h)\, p(\gamma_i)\, p(\kappa_j)\, p(\rho), \qquad (27)$$

where the product is over all K validity studies pertaining to the same constructs ξ and η (note that β_h and γ_i may vary with the studies).

The joint density (27) contains a large number of variables, prohibiting numerical integration. A modal estimation technique has to be applied to derive an estimate of the mean and variance of ρ. Leaving technical difficulties aside, we simplify for the moment by assuming a large sample normal approximation for ρ, analogously to the procedure adopted for the STSC model:

$$\rho \mid \text{data} \sim N(\tilde{\rho}, \sigma_\rho^2) \qquad (28)$$

leading to the following <u>predictive density</u> for a future validity r_f:

$$R_f \mid \text{data} \sim N(\beta_f \gamma_f \tilde{\rho}, \beta_f^2 \gamma_f^2 \sigma_\rho^2 + \frac{(1 - \beta_f^2 \gamma_f^2 \rho^2)^2}{n_f}), \qquad (29)$$

where β_f and γ_f are the instrument validities of the future predictor X_f and criterion Y_f.

Of course this model van be applied iteratively in the course of time, thus enabling a cumulative growth of knowledge on theoretical validity. The posterior density may be considered as prior density at a following stage of validation, while one or more new validity observations serve as the data. In this way an updated posterior density is obtained, which in its turn may be treated as prior density, and so on.

6. Conclusions

It appears that the procedure for validity generalization as proposed by Schmidt et al. poses a number of methodological problems: it has been employed over a much wider range than permitted by the limited, classical psychometric model on which it is, technically, based; the deductive phase of deriving the validity value to be expected in a future study is almost completely neglected; a notion of situational specificity is employed that differs from initial theorizing of e.g. Ghiselli, and that is rather narrow; rules for classifying tests, jobs and criteria are equivocal, resulting in looser classification than the underlying model would allow; the test of situational specificity has a low power and is based on inadequate logic; the validity generalization procedure is not Bayesian in character, and, when it is remodeled accordingly, it appears not to be robust to violations of the model assumptions underlying the method which are, however,

frequently present in the data.

Given these methodological problems, we feel that studies in which validity generalization in the sense of Schmidt et al. is employed, need to be reconsidered along the lines suggested in this study. There may be studies in which the test and criterion have been held constant and samples have been drawn randomly from a clearly specified population (i.e. STSC-cases). Accordingly, these offer reliable validity estimates, which are of practical importance in a clear, but limited context. An example is the study of Terborg et al. (1983) on relationships between absenteeism, job satisfaction, and commitment. But, generally, because of the methodological problems descussed in this study, the empirical limits of generalizability still have to be established. Viewing this, is may be unwise to abandon the use of detailed job analysis techniques, of the search for moderator effects (as Schmidt et al. have recommended at certain points).

It appeared that as a result of the methological problems, the outcome of the Schmidt et al. procedure may very well be biased: central tendency and lower bound validity estimates tend to be too high, and residual variances tend too be to small.

For practical applications the first consequence is more important than the second, since, generally, the central tendency and lower bound estimates obtained with the Schmidt et al. procedure are rather low. For this reason, and also

because of the simulation study presented and evaluated above, it seems likely that re-analysis along the lines presented in this study, using more representative data sets (including more specific, and objective criteria and more homogeneous job categories) would produce more distinct patterns of generalized validity, with high validity in some case and low validities in others. Such outcomes would enable a better choice of predictors and model parameters and hence contribute to the generalization procedure's total utility.

It has not been our intention to play down the importance of Schmidt, Hunter, and others work on validity generalization. We feel that they should be credited for their unique integration of ideas on sampling theory (Fisher), restiction of range (Pearson), test reliability (Spearman), summarizing research (Fisher), and Bayesian analysis (Pankoff & Roberts) into a new research method. Their creation is not only an advance in itself, it has also stirred up rigid patterns of thinking among psychologists in the field of personnel selection, which is a merit as well. Although they may not have always found a correct operationalization for their ideas rightaway, and may have overstated their opinions on certain issues, their work will undoubtedly be of influence on selection methodology in the long run.

Appendix A. The Schmidt et al. computational procedure

In the STSC approach (section 2.1), the sample consists of K attenuated and range restricted validities $r_{x_h y_i p_j}$. Since these observed validities refer to the same predictor X_h and criterion Y_i, they will be written as r_{xyp_j} ($j = 1,.., K$) for short. The variance of the validities is denoted as S^2_{obs}. Associated with each validity is a sample size n_j.

Apart from these data, distributions of predictor reliabilities, criterion reliabilities, and range restriction levels are assumed (since they are not available for every single study j). The procedure of Schmidt et al. for estimating residual validity variance and average (true) validity to be expected in future studies, proceeds as follows (we follow the description of Pearlman et al., 1980, pp. 402-406).

Estimation of the population validity

Using the n_j, a sample-size weighted mean $\bar{r}_{xyp}$ of the validity is computed. Correcting this mean by assumed average levels of criterion attenuation and restriction of range (using the corresponding assumed distributions), an estimate $\hat{\rho}_{xT_y\Pi}$ of the true, predictor attenuated validity is obtained. In some cases, predictor attenuation is also corrected for (using the mean of the assumed distribution of predictor reliabilities); then the estimate $\hat{\rho}_{T_xT_y\Pi}$ is obtained.

Estimation of residual variance

Four artifactual sources of validity variance are estimated: differences between studies in criterion reliability (variance denoted as S^2_{crit}), in predictor reliability (S^2_{pred}), in restriction of range (S^2_{rera}), and in sampling error (S^2_e).

The variance S^2_{crit} is computed as follows. The estimate $\hat{\rho}_{T_xT_y\Pi}$ is attenuated for Y again by the assumed distribution of criterion reliabilities. The variance of the resulting distribution of validities $r_{T_{xy\Pi}}$ is S^2_{crit}.

Computing S^2_{pred} starts from the mean $\hat{\rho}_{T_{xy\Pi}}$ instead of $\hat{\rho}_{T_xT_y\Pi}$, which is obtained from $\bar{r}_{xyp}$ by correcting for assumed average levels of predictor attenuation and restriction of range. Again, the point estimate $\hat{\rho}_{T_{xy\Pi}}$ is converted into a distribution of validities $r_{xy\Pi}$ by attenuating according to the assumed distribution of predictor reliabilities. The variance of this distribution is S^2_{pred}.

The variance S^2_{rera} is in turn computed from the mean observed validity $\hat{\rho}_{xy\Pi}$, which is computed from $\bar{r}_{xyp}$ by correcting for the assumed average level of restriction of range only. By the assumed distribution of range restriction levels, $\hat{\rho}_{xy\Pi}$ is transformed into a distribution of restricted validities $r_{xy\pi}$, the variance of which is S^2_{rera}.

Finally, computing S^2_e starts from each individual r_{xyp_j}. The sampling variance of r_{xyp_j} can be estimated by

(11) of section 4.1 (actually, Pearlman et al. use n_j-1 in stead of n_j). The sample-size weighted average of the K estimated sampling variances is S^2_e.

Note that computation of the four artifactual variances does not start with the same true validity estimate $\hat{\rho}_{T_xT_y\Pi}$. The procedure in fact consists of steps in which, as to say, each later step starts where the former ends. This is,according to Pearlman et al. (1980, p. 406), to ensure that the artifactual variances estimated are non-overlapping, and may therefore be added up to yield the total artifactual variance S^2_{art}:

$$S^2_{art} = S^2_{crit} + S^2_{pred} + S^2_{rera} + S^2_e.$$

The residual variance is then simply computed as $S^2_{res} = S^2_{obs} - S^2_{art}$.

Estimation of 'Bayesian prior distribution'

The estimate $\hat{\rho}_{xT_y\Pi}$ (or $\hat{\rho}_{T_xT_y\Pi}$ as the case may be), is taken as the mean of the distribution of true validities; it will be written as $\hat{\rho}$ for short. It is equal to $\bar{r}_{xyp}$ corrected for assumed average levels of criterion attenuation and range restriction. Applying the same correction factor to S^2_{res}, an estimate $\hat{\sigma}^2_\rho$ of the variance of the distribution of true validities is obtained. Assuming the latter to be normal, i.e. N ($\hat{\rho}$, $\hat{\sigma}^2_\rho$), an (e.g. 90%) lower bound estimate of the true validity is easily computed.

Appendix B. Bayesian remodeling of the STSC-model

Assuming independent priors for the parameters ρ_{xx}, ρ_{yy} and κ (and denoting these with $p(\rho_{xx})$, $p(\rho_{yy})$, $p(\kappa)$, and assuming a prior $p(\rho)$ for the central STSC parameter ρ, the joint posterior density of these parameters given the data $\{r_1, \ldots, r_j, \ldots, r_K\}$ can be obtained from the likelihood of the data given the parameters:

$$p(\rho, \rho_{xx}, \rho_{yy}, \kappa \mid r_1, \ldots, r_K) \tag{B.1}$$

$$\propto \{ \prod_{j=1}^{K} L(r_j \mid \rho, \rho_{xx}, \rho_{yy}, \kappa) \}\, p(\rho_{xx})\, p(\rho_{yy})\, p(\kappa)\, p(\rho).$$

The marginal posterior density for ρ can be obtained by integrating in (B.1) over the nuisance parameters ρ_{xx}, ρ_{yy}, and κ (e.g. using an IMSL-procedure of numerical integration).

Let

$$p^*(.) = \ln p(\rho, \rho_{xx}, \rho_{yy}, \kappa \mid r_1, \ldots, r_K),$$

then the modal equations in order to derive point estimators of the four parameters are

$$\frac{\partial p^*(.)}{\partial \rho} = \frac{\partial p^*(.)}{\partial \rho_{xx}} = \frac{\partial p^*(.)}{\partial \rho_{yy}} = \frac{\partial p^*(.)}{\partial \kappa} = 0. \tag{B.2}$$

This set of equations results in the estimators $\tilde{\underline{\mu}} = (\tilde{\rho}, \rho_{xx}, \rho_{yy}, \kappa)$. The variance-convariance matrix Σ may be approximated by the inverse of the matrix of second-order derivatives of $\rho^*(.)$ with respect to the four parameters.

Under mild conditions the posterior density (B.1) may be approximated satisfactorily by a multivariate normal

$$N(\underline{\mu}, \Sigma), \qquad (B.3)$$

(cf. Dempster et al., 1983; Lindley, 1971; Naylor & Smith, 1982). In such a case, the marginal density for ρ is the normal $N(\tilde{\rho}, \tilde{\sigma}^2_{\rho})$, where $\tilde{\sigma}^2_{\rho}$ is the first element in the diagonal of Σ, and $\tilde{\rho}$ is the modal estimator resulting from (B.2) (see (12)). Solving the modal equation (B.2) for ρ leads to (13). Because the prior densities $p(\rho_{xx})$, $p(\rho_{yy})$, and $p(\kappa)$ do not contain ρ as a parameter, (13) follows irrespective the form of these densities. Note, however, that (B.3) is an approximation; when prior densities are known, exact results can be obtained by applying a trivariate numerical integration procedure on (B.1).

To derive the predictive density of a future R_f, the true validity ρ has to be attenuated by the future r_{xx}, resulting in $\tau = \sqrt{r_{xx}}\,\rho$, where $\tau = \rho_{xTy\Pi}$. The posterior density of τ is $N(\sqrt{r_{xx}}\,\tilde{\rho}, r_{xx}\,\tilde{\sigma}^2_{\rho})$. Assuming the sampling model (cf. (11))

$$p(R_f \mid \tau) = N(\tilde{\tau}, \frac{(1-\tilde{\tau}^2)^2}{n_f}), \tag{B.4}$$

where $\tilde{\tau} = \sqrt{r_{xx}}\,\tilde{\rho}$, the predictive density of R_f is

$$p(R_f \mid r_1, \ldots, r_K) = \int_\tau p(R_f\ \tau)\ p(\tau\ r_1, \ldots, r_K)\ d\tau \tag{B.5}$$

Because (B.5) consists of normal densities, it follows (cf. Novick and Jackson, 1974) that $p(R_f\ r_1 \ldots r_K)$ is also a normal density, viz. the density given in (18).

Appendix C. Robustness of the STSC-model

With K independent studies, the likelihood of the data given the parameters $\rho_1, \ldots, \rho_K, \rho_{xx}, \rho_{yy}, \kappa$ is the product

$$L(r_1, \ldots, r_K \mid \rho_1, \ldots, \rho_K, \rho_{xx}, \rho_{yy}, \kappa) = \prod_{j=1}^{K} L(r_j \mid \rho_1, \ldots, \rho_K, \rho_{xx}, \rho_{yy}, \kappa), \tag{C.1}$$

where

$L(r_j \mid \rho_1, \ldots, \rho_K, \rho_{xx}, \rho_{yy}, \kappa) = N(g_j(\rho_j, \rho_{xx}, \rho_{yy}, \kappa), v_j)$. The function $g_j(.)$ is given by (10) with ρ_j substituted for ρ, and v_j is given by (11). As in appendix B, when independent priors are assumed, the posterior density $p(.)$ for $\rho_1, \ldots, \rho_K$ can be approximated by a multivariate $N(\tilde{\underline{\rho}}, \Sigma)$, where $\tilde{\underline{\rho}} = (\tilde{\rho}_1, \ldots, \tilde{\rho}_K)$ is estimated from the K+3 modal equations

$$\frac{\partial p^*(.)}{\partial \rho_j} = \frac{\partial p^*(.)}{\partial \rho_{xx}} = \frac{\partial p^*(.)}{\partial \rho_{yy}} = \frac{\partial p^*(.)}{\partial \kappa} = o \quad (j = 1, \ldots, K), \tag{C.2}$$

where, as in appendix B, $p^*(.) = \ln p(.)$. The variance-convariance matrix Σ is approximated by the inverse of the matrix of second-order derivatives of $p^*(.)$ with respect to the K+3 parameters.

For obtaining the predictive density of R_f, assume that c_j is the probability that the future validity study

replicates the prior study j. As in appendix B, let ρ_j be attenuated by r_{xx} to τ_j:

$$\tau_j = \sqrt{r_{xx}}\,\rho_j \qquad (C.3)$$

Assuming the future sampling model (B.4) with τ_j instead of τ and $\tilde{\tau}_j$ instead of $\tilde{\tau}$, the sampling distribution of R_f given the parameters $\tau_1, \ldots \tau_K$ is

$$p(R_f \mid \tau_1, \ldots, \tau_K) = \sum_{j=1}^{K} c_j p(R_f \mid \tau_j), \qquad (C.4)$$

i.e. a weighted average of K densities $p(R_f\ \tau_j)$. Integrating (C.4) over $\tau_1, \ldots, \tau_K$ yields the predictive density $p(R_f \mid r_1, \ldots, r_K)$ (analogously to (B.5)):

$$p(R_f \mid r_1, \ldots, r_K) = \sum_{j=1}^{K} c_j \int_{\tau_1} \cdots \int_{\tau_K} p(R_f \mid \tau_1, \ldots, \tau_K)$$

$$p(\tau_1, \ldots, \tau_K \mid r_1, \ldots, r_K)\, d\tau_1, \ldots d\tau_K. \qquad (C.5)$$

Approximating the posterior $p(\rho_1, \ldots, \rho_K \mid r_1, \ldots, r_K)$ by

the multivariate normal $N(\tilde{\rho}, \Sigma)$, (C.3) implies that the posterior $p(\tau_1, \ldots, \tau_K \mid r_1, \ldots, r_K)$ is also multivariate normal. Since a multivariate normal can be written as the product of a normal marginal and a normal conditional density:

$$p(\tau_1, \ldots, \tau_K \mid r_1, \ldots, r_K) = p(\tau_1 \mid r_1, \ldots, r_K)\, p(\tau_2, \ldots, \tau_K \mid \tau_1, r_1, \ldots, r_K),$$

it follows that the integration in (C.5) yields the result (20).

References

Algera, J.A., Jansen, P.G.W., Roe, R.A. & Vijn, P. (1984)

Validity generalization: some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology*, 57, 197-210.

Arvey, R.D., Maxwell, S.E. & Mossholder, K.M. (1979)

Even more ideas about methodologies for determining job differences and similarities. *Personnel Psychology*, 32, 529-538.

Baehr, M.E. (1967)

A factorial framework for job descriptions for higher-level personnel. Chicago: Industrial Relations Center, University of Chicago.

Bobko, P. & Rieck, A. (1980)

Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement*, 4, 385-398.

Brown, S.H. (1981)

Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology*, 66, 664-670.

Dowell, B.E. & Wexley, K.N. (1978)

Development of a work behavior taxonomy for first-line supervisors. *Journal of Applied Psychology*, 63, 563-572.

Drasgow, F. & Miller, H.E. (1982)

Psychometric and substantive issues in scale construction and validation. *Journal of Applied Psychology*, 67, 268-279.

Dunnette, M.D. (1963)

A note on *the* criterion. *Journal of Applied Psychology*, 47, 251-254.

Dunnette, M.D. (1972)

Validity study results for jobs relevant to the petroleum refining industry. Washington D.C.: American Petroleum Institute.

Efron, B. & Morris, C. (1981)

Stein's paradox in statistics. *Scientific American*, 76, 341-353.

Ekehammar, B. (1974)

Interactionism in personality from a historical perspective. *Psychological Bulletin*, 81, 1026-1048.

Elshout, J.J., Overbeek, H. van, Roe, R.A. & Vijn, P. (1979)

Testing the hypothesis that rho = 0 in selected samples (case I). *Educational and Psychological Measurement*, 39, 573-576.

Everitt, B.S. & Hand, D.J. (1981)

Finite mixture distributions. London: Chapman and Hall.

Ewart, E.S., Seashore, S.E. & Tiffin, J. (1941)

A factor analysis of an industrial merit rating scale. *Journal of Applied Psychology*, 25, 481-486.

Fleishman, E.A. & Ornstein, G.E. (1960)

An analysis of pilot flying performance in terms of component abilities. *Journal of Applied Psychology*, 44, 146-155.

Fogli, L., Hulin, C.L. & Blood, M.R. (1971)

Development of first-level behavioral job criteria. *Journal of Applied Psychology*, 55, 3-8.

Forehand, G.A. (1963)

Assessments of innovative behavior: Partial criteria for the assessment of executive performance. *Journal of Applied Psychology*, 47, 206-213.

Forsyth, R.A. (1971)

An empirical note on correlation coefficients corrected for restriction in range. *Educational and Psychological Measurement*, 31, 115-123.

Forsyth, R.A. & Feldt, L.S. (1969)

An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *Educational and Psychological Measurement*, 29, 61-72.

Ghiselli, E.E. (1959)

The generalization of validity. *Personnel Psychology*, 12, 397-402.

Ghiselli, E.E. (1966)

The validity of occupational aptitude tests. New York: Wiley.

Ghiselli, E.E. (1973).

The validity of aptitude tests in personnel selection. *Personnel Psychology*, 26, 461-477.

Grant, D.L. (1955)

A factor analysis of managers' ratings. *Journal of Applied Psychology*, 39, 283-286.

Greener, J.M. & Osburn, H.G. (1979)

An empirical study of the accuracy of corrections for restriction in range due to explicit selection. *Applied Psychology Measurement*, 3, 31-41.

Gross, A.L. & Perry, Ph. (1983)

Validating a selection test, a predictive probability approach. *Psychometrika*, 48, 1, 113-127.

Guion, R.M. (1961)

Criterion measurement and personnel judgments. *Personnel Psychology*, 14, 141-149.

Guion, R.M. (1965)

Personnel testing. New York: McGraw-Hill.

Guion, R.M. (1976)

Recruiting, selection, and job placement. In: Dunnette, M.D. (ed.), *Handbook of Industrial and Organizational Psychology*, 777-828. Chicago: Rand McNally.

Hays, W.L. (1973)

Statistics for the social sciences. New York: Holt, Rinehart & Winston.

Hemphill, J.K. (1959)

Job descriptions for executives. *Harvard Business Review*, 37, 55-67.

Hemphill, J.K. (1960).

Dimensions of executive positions. Research Monographs, no 98. Ohio State Univeristy: Bureau of Business Research.

Hunter, J.E., Schmidt, F.L. & Pearlman, K. (1982)

History and accuracy of validity generalization equations: A response to the Callender and Osburn reply. *Journal of Applied Psychology*, 67, 853-858.

James, L.R. (1973)

Criterion models and construct validity for criteria. *Psychological Bulletin*, 80, 75-83.

Jansen, P.G.W., Roe, R.A., Vijn, P. & Algera, J.A. (1984)

Validity generalization: Critique and proposals. Internal report.

Jöreskog, K.G. (1973)

A general method for estimating a linear structural equation system. In : Goldberger, A.S. & Duncan, O.D. (eds.), *Structural equation models in the social sciences.* New York: Seminar Press, 85-112.

Jöreskog, K.G. (1974)

Analyzing psychological data by structural analysis of covariance matrices. In: Atkinson, R.C., Krantz, D.H. & Suppes, P. (eds.), *Contemporary developments in mathematical psychology* (Vol. II). San Francisco: Freeman & Col., 1-56.

Jöreskog, K.G. (1978)

Structural analysis of covariance and correlation matrices. *Psychometrika*, 43, 443-477.

Jöreskog, K.G. & Sörbom, D. (1978)

LISREL IV. Analysis of linear structural relationships by the method of maximum likelihood. Chicago: International Educational Services.

Kirchner, W.K. (1966)

Relationships between supervisory and subordinate ratings of technical personnel. *Journal of Industrial Psychology*, 3, 57-60.

Kraemer, H.C. (1979)

Tests of homogeneity of independent correlation coefficients. *Psychometrika*, 44, 329-335.

Lawshe, C.H. (1952)

Employee selection. *Personnel Psychology*, 5, 31-34.

Lawshe, C.H. & Balma, M.J. (1966)

Principles of personnel testing. Second edition. New York: McGraw Hill.

Lee, R., Miller, K.J. & Graham, W.K. (1982)

Corrections for restriction of range and attenuation in criterion-related validation studies. *Journal of Applied Psychology*, 67, 637-639.

Lindley, D.V. (1971)

Bayesian statistics, a review. *Regional Conference Series in Applied Mathematics*, S.I.A.M.

Lindley, D.V. & Smith, A.F.M. (1972)

Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society*, 34, 1-41.

Linn, R.L., Harnisch, D.L. & Dunbar, St.B. (1981)

Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology*, 66, 655-663.

Lissitz, R.W., Mendoza, J.L., Huberty, C.J. & Markos, H.V. (1979).

Some further ideas on a methodology for determining job similarities/differences. *Personnel Psychology*, 32, 517-528.

Lord, F.M. & Novick, M.R. (1968)

Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Lumsden, J. (1976)

Test theory. *Annual Review of Psychology*, 27, 251-280.

McCormick, E.J. (1976)

Job and task analysis. In: Dunnette, M.D. (ed.), *Handbook of Industrial and Organizational Psychology*, 651-696. Chicago: Rand McNally.

Magnusson, D. (ed.) (1981)

Toward a psychology of situations. An interactional perspective. Hillsdale: Lawrence/Erlbaum.

Monahan, C.J. & Muchinsky, P.M. (1983)

Three decades of personnel selection research: A state-of-the-art analysis and evaluation. *Journal of Occupational Psychology*, 56, 215-225.

Naylor, J.C. & Smith, A.F.M. (1982)

Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31, 214-225.

Novick, M.R. & Jackson, P.H. (1974)

Statistical methods for educational and psychological research. New York: McGraw-Hill.

Olson, C.A. & Becker, B.E. (1983)

A proposed technique for the treatment of restriction in range in selection validation. *Psychological Bulletin*, 93, 137-148.

Osburn, H.G., Callender, J.C., Greener, J.M. & Ashworth, S. (1983)

Statistical power of tests of the situational specificity hypothesis in validity generalization studies: a cautionary note. *Journal of Applied Psychology*, 68, 1, 115-122.

Overbeek, H.J. van (1974)

Een onderzoek naar 'restriction of range' bij RHO = 0. Universiteit van Amsterdam: Psychologisch Laboratorium.

Pearlman, K., Schmidt, F.L. & Hunter, J.E. (1980)

Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, 65, 373-406.

Peres, S.H. (1962)

Performance dimensions of supervisory positions. *Personnel Psychology*, 15, 405-410.

Prien, E.P. (1965)

Development of a clerical position description questionnaire. *Personnel Psychology*, 18, 91-98.

Roberts, K.H. & Glick, W. (1981)

The job characteristics approach to task design: a critical review. *Journal of Applied Psychology*, 66, 2, 193-217.

Roe, R.A. (1979)

The correction for restriction in range and the difference between intended and actual selection. *Educational and Psychological Measurement*, 39, 551-559.

Roe, R.A. (1983)

Grondslagen der personeelsselektie. Assen: Van Gorcum.

Roe, R.A. (1984)

Advances in performance modeling: the case of validity generalization. *Paper presented at the Symposium 'Advances in Testing'*. International Test Commission. Acapulco, Mexico, Sept. 6.

Roe, R.A. & Elshout, J.J. (1972)

Some new formulas for the correction for restriction of range. *Nederlands Tijdschrift voor de Psychologie*, 27, 134-139.

Roe, R.A., Algera, J.A., Jansen, P.G.W. & Vijn, P. (1983a)

Ernst met methodische deskundigheid. *De Psycholoog*, 18. 133-142.

Roe, R.A., Algera, J.A., Jansen, P.G.W. & Vijn, P. (1983b)

De olifant en de nieuwe kleren van de keizer: een antwoord aan Hofstee. *De Psycholoog*, 18, 503-512.

Ronan, W.W. (1963)

A factor analysis of eight job performance measures. *Journal of Industrial Psychology*, 1, 107-112.

Ronan, W.W. & Prien, E.P. (eds.) (1971)

Perspectives on the measurement of human performance. New York: Appleton Century Crofts.

Rubin, D.R. (1980)

Using empirical Bayes techniques in the law school validity studies. *The Journal of the American Statistical Association*, 75, 801-827.

Rubin, D.R. (1981)

Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 4, 377-400.

Rusch, C.H. Jr. (1953)

A factorial study of sales criteria. *Personnel Psychology*, 6, 9-24.

Schmidt, F.L., Gast-Rosenberg, I. & Hunter, J.F. (1980)

Validity generalization results for computer programmers. *Journal of Applied Psychology*, 65, 643-661.

Schmidt, F.L. & Hunter, J.E. (1977)

Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.

Schmidt, F.L. & Hunter, J.E. (1981)

Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128-1137.

Schmidt, F.L., Hunter, J.E. & Caplan, J.R. (1981a)

Validity generalization results for two job groups in the petroleum industry. *Journal of Applied Psychology*, 66, 262-273.

Schmidt, F.L., Hunter, J.E. & Pearlman, K. (1981b)

Task differences of aptitude test validity in selection: a red herring. *Journal of Applied Psychology*, 66, 166-185.

Schmidt, F.L., Hunter, J.E. & Pearlman, K. (1982)

Progress in validity generalization: comments on Callender and Osburn and further developments. *Journal of Applied Psychology*. 67, 835-845.

Schmidt, F.L., Hunter, J.E., Pearlman, K. & Shane, G.S. (1979)

Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, 32, 257-281.

Schmidt, F.L. & Kaplan, L.B. (1971)

Composite vs. multiple criteria: a review and resolution of the controversy. *Personnel Psychology*, 24, 419-434.

Schmitt, N., Gooding, R.Z., Noe, R.A. & Kirsch, M. (1984)

Meta analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407-422.

Schultz, D.G. & Siegel, A.I. (1964)

The analysis of job performance by multidimensional scaling techniques. *Journal of Applied Psychology*, 48, 329-335.

Seashore, S.E., Indik, B.P. & Georgopoulos, B.S. (1960)

Relationships among criteria of job performance. *Journal of Applied Psychology*, 44, 195-202.

Siegel, A.I. & Pfeiffer, M.G. (1965)

Factorial congruence in criterion development. *Personnel Psychology*, 18, 267-280.

Smith, P.C. (1976)

Behaviors, results and organizational effectiveness: the problem of criteria. In: Dunnette, M.D. (ed.), *Handbook of Industrial and Organizational Psychology*, 745-775. Chicago: Rand McNally.

Smith, P.C. & Kendall, C.M. (1963)

Retranslation of expectations: an approach to the construction of unambigious anchors for rating scales. *Journal of Applied Psychology*, 47, 149-155.

Stark, S. (1959)

Research criteria of executive success. *Journal of Business*, 32, 1-14.

Toops, H.A. (1944)

The criteria. *Educational and Psychological Measurement*, 4, 271-297.

Tornow, W.W. & Pinto, P.R. (1976)

The development of a managerial job taxonomy: a system for describing, classifying and evaluating executive positions. *Journal of Applied Psychology*, 61, 410-418.

Vernon, P.E. (1964)

Personality assessment: A critical survey. London: Methuen.

Viana, M.A.G. (1980)

Statistical methods for summarizing independent correlational results. *Journal of Educational Statistics*, 5, 83-104.

Vijn, P. (1983)

Prior information in linear models. (Unpublished doctoral dissertation.) University of Groningen.

Wiley, L. (1964)

Relation of characteristics ratings to performance ratings. *Journal of Industrial Psychology*, 2, 7-15.

Winne, Ph.H. & Belfroy, M.J. (1982)

Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, 19, 125-134.

Wofford, J.C. (1970)

Factor analysis of managerial behavior variables. *Journal of Applied Psychology*, 54, 169-174.

Wolff, Ch.J. de (1970)

Beoordelingen als criteria. In: Drenth, P.J.D., Willems, P.J. & Wolff, Ch.J. de (red.), *Bedrijfspsychologie, onderzoek en evaluatie*. Kluwer/Van Loghum Slaterus, 677-694.

158083

866592